We thank the reviewers for their thoughtful comments, which have very much improved our paper! Please find attached a “tracked changes” version of the manuscript.

R1

Review of Towards robust community assessments of the Earth’s climate sensitivity

General Comments
The authors present a nice overview of how Bayesian statistics can be used to make an assessment on climate sensitivity. Importantly, they discuss the various choices taken to make inference on climate sensitivity and how the choices affect the resulting estimate. While the paper is written well and most information is presented well, I feel the key concepts are somewhat obscured by their notation and lack of clarity. Below, I include some specific examples that I feel could be improved upon.

Specific Comments

Numerous terms are undefined. For example, in Eq. (1), what is M0? (It is not defined until two pages later). Same with Mα and other models. Also, while it may be colloquial for some, F2×CO2 and ΔT2×CO2 are undefined.

We agree! We’ve now defined these terms.
Various mathematical terms are undefined.

• Line 105 - What is N(−8.43,2)? I assume you mean a normal distribution with mean −8.43 and variance 2. Also, math cal font is used for N in section 7 and not for line 105.

Fixed. As an aside, thank you for calling this out. Often in the literature it’s unclear whether s is the standard deviation or the variance in the notation N(m,s).

• Line 201 - What is ΔT′(x) and what is x?

This has been removed.

In multiple areas the authors discuss the idea of reducing uncertainties but do not explicitly say which uncertainties are being (or will be) reduced. I think this is a common misnomer when discussing statistical concepts. Some uncertainties are irreducible, such as types of data uncertainty, and some are reducible, such as types of model uncertainty. A key concept in statistics is being able to identify each type and providing the best possible quantification of each - e.g. an appropriate quan- tification of irreducible uncertainty and reducing all other uncertainty if possible. I feel strongly that the manuscript would benefit if this distinction were made clear.

We've now tried to make this much clearer. This paper is intended as a guide for those embarking on large-scale expert assessments, whether a follow-up to Sherwood et al 2020 or assessments of other quantities like TCR or ZEC. To that end, we're focusing on identifying the areas where unavoidable subjective decisions (i.e., expert judgment) enter the analysis.

Line 60 - "... update our prior beliefs P(Θ)... " Are you making the distinction that you can update you P(Θ|Y) if new information becomes available? If so, I think this needs to be reworded. Otherwise, I do not believe Bayes' Theorem says update our prior beliefs. Instead, if you have a prior belief or knowledge, P(Θ), you can get an estimate of the probability of Θ given data/evidence Y using that prior knowledge. If more data becomes available, you could then refine that belief and use a new prior. This is distinct from updating your prior belief.

Yes, this is a tough one to convey. We were trying to balance technical rigor with comprehensibility for a wider audience. Please see the rewritten "Analysis Framework" section:

"This framework allows us to use our prior understanding of the parameter values to calculate the posterior probabilities P (Θ|Y, M) of the model parameters given the evidence. This posterior can be updated as new evidence becomes available."

Section 2 Analysis framework - This is a crucial section for your paper and I would like to see it expanded. Throughout the rest of the paper, terms like posterior, marginal, joint probability density, ... are used but not defined. The general reader of ESD may be unfamiliar with these terms. The latter parts of the paper would be easier to follow if these terms are defined in the context of section 2. Additionally, how does one get P (Θ|Y)? Is it as simple as writing distributions down and using Bayes' Theorem? What if the distribution is not tractable, how would that be handled? Expanding on some of the steps needed to make inference on Θ within
this section will help orient readers as to why this is such a difficult and important problem, and how they can take what you have shown and apply it to their own analysis.

We've substantially rewritten the entire section for clarity- see the revised text below. The section now begins with a definition of evidence, model, and prior. (Note that now we explicitly represent the prior dependence on the model). We then define prior, likelihood, and posterior as relevant for the problem of assessing ECS.

"Bayes' Theorem can be written as

$$P(\Theta|Y, M) = P(Y|\Theta, M))P(\Theta|M)/P(Y|M) . \quad (3)$$

Here, we will define these terms as they apply to the problem of estimating climate sensitivity.

Evidence The evidence Y used to constrain climate sensitivity consists of the global mean

temperature change ΔT in response to a forcing ΔF as well as, in non-equilibrium

states, the net energy imbalance ΔN . We have estimates of these quantities for the

historical period (derived from observations and models) and for past climate states

(derived from paleoclimate proxies and models), and Y therefore consists of multi-

ple lines Y1 . . . Yn. For example, S20 used process-based understanding of underlying

physics, recent observations, and proxy-based reconstructions of past climates to assess

S.

The model M codifies how we interpret the evidence Y . It specifies the parameters

Θ whose posterior distributions we estimate. For example, in the simple energy balance

model denoted M0, there is only one parameter and $\Theta = \lambda$. The model determines the

likelihood $P(Y|\Theta, M)$ of observing the data given particular values of the parameters

Θ. We discuss methods for calculating this likelihood in Section 4.1.

The prior probability distribution $P(\Theta|M)$ reflects prior beliefs or knowledge about

the model parameters Θ. For example, in the simple model M0, the community assess-

ment S20 adopted a uniform prior on λ as a baseline choice, choosing not to rule out

net positive feedbacks (and therefore an unstable climate) a priori. “

Line 120 - The notation surrounding this equation is confusing. It appears as though you are treating $Y = (\Delta T, \Delta F)$ as normally distributed random variables where the mean and standard deviation of each are estimated from experts. You define the joint probability density of ΔT and ΔF as $J(Y) \equiv J(\Delta T, \Delta F)$. You then marginalize over ΔT and ΔF and somehow get a likelihood of the evidence as $P(\Delta T, \Delta F|\lambda)$. However, this equation (A.3), does not contain either ΔT or ΔF because they have been marginalized out. Instead it contains their mean and standard deviation (assumed fixed?) that are estimated by experts. My confusion is in your definition of evidence and how ΔT and ΔF (or their

mean and standard deviation) play a role in that evidence. I think this could be fixed by being more clear on your notation and the steps taken to arrive at the equation on line 120 (and subsequent equations).

You're right that this section was lacking clarity. In climate sensitivity assessments, the evidence is necessarily uncertain. We don't have point measurements of temperature, forcing, etc. Instead, we can assess the literature and come up with a joint probability density ρ(ΔT, ΔF) (here, this is the product of the PDFs for ΔT and ΔF, but more complex distributions reflecting correlated errors are possible). In this paper, we make the argument that given a model M described by a curve C in evidence space, the likelihood can be approximated by calculating the probability mass along the curve. S20 calculated historical and paleo likelihoods using a similar method but different language; to our knowledge no other climate statistics paper has employed the probability mass method. We've rewritten the section to be clearer.

Model Mα - This is slightly confusing to me. By definition, α = ∂λ/∂ΔT is a function of λ. However, you assign independent priors to λ and α when α is constrained by the value of λ. Is there justification for specifying independent priors here? Or is this done for illustrative purposes? If so, I feel it is important to note they are not independent.

In this case, α is a constant, not a continuous function of λ. There is no a priori reason why the uncertainty in the net feedbacks should be correlated with the uncertainty in their rate of change with global mean temperature. The point here is to make N a quadratic function of ΔT, N = ΔF + aΔT +bΔT^2 where a = λ and b = α/2.

Eqn. (7) - Same as above comment except now for Δλ.

Δλ is a Gaussian that we specify, which is independent of λ _hist. Obviously the fact that the net feedback λ = λ _hist + Δλ means that λ is not independent of λ _hist or Δλ? Perhaps this is where the confusion arises?

Section 7.1 - I rather like this section and I think it puts a lot of the paper into context. However, I feel as though some terms are not defined and potentially un- known to the general ESD reader. A (Bayesian) hierarchical model is left undefined and for the reader to interpret. Generally, a BHM is defined in terms of data, pa- rameter, and sometimes hyperparameter models. It might help contextualize your message if you state what a hierarchical model is in terms defined from section 2 and then connect it to equations (9)-(12).

We've rewritten the section for improved clarity, and no longer provide a meta-analysis estimate of radiative forcing, which we now realize enlarged the scope of our paper unnecessarily. To appeal to a broader audience, we've also de-emphasized the

hierarchical nature of Bayesian meta-analysis, stressing the physical meaning of the priors on the hyperparameters mu and tau. This has also helped us make clearer recommendations: these priors must be specified, and we suggest that organizers of future assessments do so (as opposed to querying the broader community of experts).

Line 261 - This is a very bold claim. I would argue it is one useful application of hierarchical modeling, but maybe not one of the most.

Removed.

Line 396 - C is already defined as the curve.

fixed

Technical Comments

1. A few citations have typos in or surrounding how they are placed in the text 2. Figure 1 - needs labels

fixed.

3. Line 9 - multiple twice fixed
4. Line 234 - Missing) This discussion has been removed from the paper.

5. Line 241 - "... that this is are the ..." 6. Line 243 - Missing) This discussion has been removed from the paper.

R2

This manuscript presents an assessment of the uncertainty associated with Bayesian inferences of the effective climate sensitivity parameter, S as assessed by Sherwood et al. (2020). The authors clearly point out three sources of uncertainty that were not previously addressed in Sherwood et al. (2020): evidence uncertainty, structural uncertainty, and prior uncertainty and illustrate each with examples related to the Last Glacial Maximum. They conclude with a recommendation of how to combine multiple lines of evidence to constrain S that will allow for rapid updates in light of new evidence in the future.

The manuscript creates more awareness in terms of the uncertainty involved in Bayesian inferences such as that of Sherwood et al. (2020). The topic is both

interesting and important. The authors provided several interesting examples, however, in my opinion, the manuscript reads somewhat esoteric for the atmospheric science community. Overall, I have a few suggestions that are quite minor in nature, mostly regarding clarifying the main messages for individual sections for the authors to consider before recommending publication.

For context, it would be helpful if the authors could please specify in the Introduction the uncertainty that was discussed in Sherwood et al. (2020) and then follow this with the additional complementary detail that they consider.

This is a helpful comment. We've now tried to clarify the aims of our paper vs those of S20 (and IPCC AR6) and have added the text below.

IPCC AR6 assessed confidence in the range of S based on suport from individual lines of evidence, and the medium confidence assessed was in large part due to the fact that not all lines of evidence supported the same upper bound. By contrast, S20 sought to provide a robust estimate by combining lines of evidence in a coherent Bayesian framework. However, S20 used baseline priors and estimates of the evidence and investigated the impact of alternate choices as sensitivity tests rather than attempt to combine multiple priors, estimates, and expert judgements into a single posterior probability distribution. In both IPCC AR6 and S20, as in almost all previous assessments, the means by which disagreements among experts were resolved or handled was not necessarily made transparent. This paper presents some lessons learned by two authors of S20 and attempts to chart a way forward.

Our goal is not to provide a single updated estimate but rather to understand where unavoidable subjective decisions enter in to the analysis and to present a framework for systematically and fairly incorporating the subjective judgements of multiple experts.

Section 6.1: this section on comparing "apples to apples" when considering different lines of evidence to constrain S is an important point that was raised, and the example is interesting, however, the main point is unclear and the description is too roundabout. In this particular example, are the authors recommending the Bayes factor as a solution for evaluating the reliability of different lines of evidence? I would suggest a clear statement at the end of this section regarding the authors' recommendation regarding how to treat the issues of fairly comparing different lines of evidence when constraining S.

This section was initially quite confusing, and we apologize. We've rewritten it to stress the importance of the interpretive model in rendering multiple lines of evidence compatible and resolving the "Twin Peaks" problem. The "best" model will depend on prior knowledge: if we have reason to believe that one interpretive model is better than another, then the Twin Peaks problem may not be an actual problem: a small area of

overlap between posteriors updated with multiple lines of evidence may then constrain the parameters extremely well. However, if we have no reason to believe that one model is “better” than any other, then models that render lines of evidence more compatible will be preferred.

Lines 248-251: This brings up the important point that subjectivity is an issue, however, other than suggesting more transparency in terms of making subjective decisions, it does not seem that the authors are reducing any subjectivity. Please clarify. For example, on lines 319-320, why use a weighted average? This choice itself is apparently subjective. A well-justified recommendation could better convince others to follow these recommendations in the future.

We’ve clarified to explain that we’re not necessarily reducing subjectivity: as Figure 1 shows, there are unavoidable subjective decisions in every analysis. Instead, we’re arguing that these decisions need to be clearly communicated, and that expert judgment should be clearly specified in the form of priors. We’ve now added specific “recommendation” sections throughout Section 7 to make this clearer. In short, we present a method for achieving transparency and clarity on necessary subjective decisions made.

I recommend that the authors also summarize their specific recommendations for future Bayesian analysis in the Abstract. There is space for it in the Abstract.

Good suggestion, done.

Typographic errors

Lines 242: “that this is are the model…” should be “that this is the model…”
This discussion has been removed from the paper.
Line 269: “about a some true…” should be “about some true…” fixed
Line 279: “As an specific” should be “As a specific” fixed
Line 329: Latex compilation error for Table fixed, table moved to supplementary material
Line 343: “distributions” should be singular fixed
Line 337: Isn’t Expert B also open to the idea that lambda can be positive too in Figure 8a?
For consistency and convenience, please number all equations, including the ones on page 13, even if not explicitly referenced in the text. fixed
The notation N(x,y) was defined in Sherwood et al. (2020) but not in this manuscript. Please define it in this manuscript. defined.

R3

In this manuscript the authors discuss issues and possible ways forward with bayesian based assessments of climate sensitivity. This follows on a first major attempt by Sherwood et al. (2020), which was influential on the IPCC report (Forster et al. 2021). I have no major issues with the manuscript, and I think it is great that the approach and issues are discussed openly. I would have hoped, before reading it, that the text would have been even more accessible to a wider audience, but in several places there is quite a bit of statistics jargon.The minor recommendations below are not fully addressing this issue, and I would leave it up to the authors to consider this issue. Anyway, I see no major obstacles to publication.

2-3, Most of these lines of evidence constrain S or ECS, not feedback in isolation.

We've replace this with "The uncertainty inS primarily results from uncertainties in net physical climate feedback, usually denoted as \lambda. "

26-27, Note that IPCC did not use the bayesian method, so perhaps state what they did. Furthermore, the medium confidence is due to not all individual lines of evidence supporting a 95th percentile close to 5 K.

We've now noted this in the Introduction- but note that assessing confidence based on support from individual lines of evidence is at odds with our proposed method of combining lines of evidence in a coherent Bayesian framework.

26, the reference should be Forster et al. (2021) fixed

53, remove one instance of 'is'

Removed

59, why use quotation marks here?

Removed

63-64, I am not sure this was what the lines of evidence were called in S20.

It wasn't, but we've adopted this notation for clarity. Note the rewritten Analysis section that now clearly defines all terms.

96-97, See also Annan et al. (2022) for a discussion of how T20 might be cold-biased and over-confident due to reliance on a single-model prior.

Agreed. We've now noted this and emphasized that the two studies are not comparable, and use them only to illustrate the impact of evidence uncertainty

98, These are not simply two equally valid or comparable studies. S20 is an assessment in which, in principle, the authors took into account a much broader evidence base than used by T20.

Agreed- see response above.

105, should probably be '-9.6'

Fixed

125, missing closing parenthesis.

Fixed

Section 4, Perhaps check out https://doi.org/10.5194/cp-19-323-2023.

Cited.

136-137, The quadratic model is that ECS changes monotonically until an instability occurs. There is some evidence that ECS will increase with warming, but we also know there were snow ball Earth instabilities in the past, so ECS must increase into a bifurcation at cooling temperatures. The shape of this function is not well known, other than that there is a minimum not too far from our current climate. What I am getting at is that the quadratic model is only half the story, and the evidence for a positive alpha comes mostly from warmer climates. It might be negative at colder temperatures, which effectively means the model is no longer valid.

This is of course possible, and we've now alluded to the potential bifurcation of alpha with temperature in our list of potential models.

166-167, A bit of an understatement. Since we are here it is not physically possible that climate sensitivity is negative, i.e. the system is unstable, so prior knowledge forbids negative climate sensitivity.

In S20, flat uniform priors were placed on the individual feedback components in the process evidence, implying a prior on the net feedback that puts equal weight on positive and negative values. Clearly, as you point out, this is unphysical. S20 dealt with this by removing low-likelihood values of individual feedback components, while here we prefer to allow unstable climates to be definitively ruled out by the evidence. In the prior uncertainty section we do use the knowledge derived from the process evidence as an alternate prior.

Figure 3, I think this would have been useful earlier, just a thought.

Thanks! We've now moved it to Figure 1.

198, + Modak and Mauritsen https://doi.org/10.5194/acp-23-7535-2023

Cited.

203, Note that this is based on the AMIPII dataset which produces the largest pattern effect of all SST reconstructions (https://doi.org/10.5194/acp-23-7535-2023)

Yes, we've now noted that the prior on Delta lambda used in S20 may be too weighted toward large values (and possibly too narrow), but we use the S20 marginal historical likelihood for illustrative purposes.

223-224, I find this argument weak: it looks better, hence it must be right? On could also state that we trust historical warming, and we will use alpha as a fudge factor on the LGM evidence to make it match the historical.

We've substantially rewritten this section, but we do feel that the "Twin Peaks" problem is worth noting. If posteriors updated by multiple lines of evidence have a small region of overlap, one of two things is true: either we can be highly confident in the resulting estimates, or the lines of evidence are not, in fact, measuring the same thing. In the absence of prior knowledge regarding either of these possibilities, a model that brings posteriors from different lines into better agreement has more evidence to support it. If we do have prior knowledge or beliefs that we are indeed comparing "apples to apples", this is reflected in the term P(M1)/P(M2), and causes the resulting posterior derived from multiple lines of evidence to be more sharply peaked.

Equation above 235 and the equation shortly thereafter, there is a missing closing parenthesis.

This has been removed.

244 (some issue with line numbering, that is the line just above 245), A strange formulation, I would say "If using T20 evidence, more agreement with historical evidence is obtained if assuming alpha is close to zero." If one were to use Annan et al. (2022) or a weaker pattern effect estimate, then the result would be different.

We agree completely with this statement. We've now rewritten the text to say "Clearly, the ``best" model depends on the evidence used, the prior knowledge of whether we are comparing ``apples to apples", and the priors we place on λ, $\Delta \lambda$, and α."

288, Perhaps comment why this entire range is warmer than the range estimated above? It is the same evidence, I suppose, so a nice example of how a too wide asymmetric prior can bias the posterior.

We've added "and warmer"- note that this is because a "fixed effects" model will simply treat cool estimates as outliers, hence the warmer values.

339, missing table number

Table deleted in rewrite.

Table 3, '1,9' -> '1.9'

Table deleted in rewrite.

342, The authors are careful to write she/he in other places, but not here. Why is it that the extremely overconfident expert is male?

We've added she pronouns.

Section 8, I felt this section was hard to read, and I feel like it could be shortened and sharper.

We've made some big changes, not least identifying specific recommendations.

353, "where do those estimates or measurements come from", I think this text is open to too much interpretation.

We've replaced this with "how do we decide what counts as "evidence", which we hope is clearer.

356, please state which section.

This has been rewritten to describe the overall strategy.

371, delete one instance of 'make'
Done.

Comment by John Eyre

General Comments

1. This is an interesting paper, containing both results and discussion of method that are likely to be helpful to the community involved in assessments of climate sensitivity.

Thank you.

2. I suggest that the paper could be improved substantially by adopting different terminology: by transferring from a terminology appropriate to a subjective interpretation of probability theory (flowing from a subjectivist theory of knowledge) to a terminology appropriate to objective theories of both probability and knowledge. This would involve no changes to the equations or the results, as the mathematics of the probability theory would be unchanged, but it would change the way in which the mathematics is interpreted in terms of its relation to the real world.

In Bayesian statistics there are multiple schools of thought, including subjectivist Bayesianism and Objectivist Bayesianism (see Gelman and Hennig 2017 for a review). There is a lack of consensus on the best terminology to use in Bayesian statistics, with many different approaches being advocated by different researchers. Additionally, Gelman and Hennig (2017) argue that the words 'objective' and 'subjective' in statistics discourse are used in a mostly unhelpful way. Many of your points certainly might be valid, but we are not experts on the philosophy of statistics, and the purpose in this study is not to address longstanding debates about the terminology of Bayesian statistics. It is to propose improvements to the way that Bayesian statistics can be applied to the problem of organizing community assessments of evidence.

Our understanding is that subjectivist language is more commonly used than objectivist language in the literature and; the terminology we have adopted is commonly used if not universally agreed upon. Perhaps more importantly, the colloquial understanding of Bayesian methods (such as it exists) regards Bayesian inference as analogous to a learning process, in which prior beliefs are replaced by updated beliefs in light of evidence. That said, we find your arguments compelling and will also adopt some of your specific proposals - please see below for details.

3. Specifically, I suggest the term "belief" (particularly in the term "prior belief") be changed throughout. In most place it could be replaced by "estimate" or "information" or "knowledge". In other places the meaning is different, and it would be better replaced by "assumption". Similarly, I suggest that the term "subjective" is over-used. In most places, what is described as "subjective" is in fact objective, i.e. it is inter-subjectively shared and criticised. In most cases, this sharing and criticism is of the very high standard expected of publications in the scientific literature.

We've replaced most instances of "belief" with "belief and/or knowledge" or just "prior". However, in some cases, we humbly suggest that we do mean "belief" in the actual sense. See, e.g., our discussion of how to handle experts with overly narrow and/or biased priors. While scientific knowledge should, in theory, be updated systematically and dispassionately with evidence, it is not our experience in working with actual scientists that this always happens. Scientists, as humans, approach questions with priors informed not just by previous evidence but emotional states, ego, cultural background, political biases, etc. Our hypothetical scientist C strongly believes climate sensitivity to be low, not necessarily because s/he has extensive knowledge others with broader priors do not, but because s/he wants it to be.

4. So, if accepted, these comments would imply numerous changes to the text, but ones that could be made without changes to the structure and scientific content of the paper.

5. A subjective theory of knowledge was widely accepted up to the middle of the 20 th century. It was accompanied by a subjective interpretation of probability theory in general
and of Bayes theorem in particularly. This interpretation was heavily criticised by Karl

Popper in many of his key works. The preference for an objective rather than a subjective,
theory of probability is discussed most cogently by Popper in “Realism and the aim of science” (1983). Chapter 1 of Part II is entitled “Objective and subjective probability”, and
the comments in this review are intended to be consistent with Popper’s treatment of these
problems. In summarising the difference between these two approaches to probability theory, Popper says (section 7, para 1): “... The subjectivist takes a as his hypothesis and
P(a|b) as our degree of belief in it, whilst the objectivist takes ’P(a|b)=r’ as his hypothesis.
(He may or may not believe in it.) ...” . (The subjectivist example here stands for the probability that hypothesis a is true given evidence b, but it applies equally to the case where
a is the estimate of a quantity and b is the observational evidence supporting it.)

If one accepts Popper’s criticisms, then the subjective interpretation of probability is both out-moded and unnecessary (although it appears to linger on in some text books on philosophy of science and on statistics).

We’re far from experts in the philosophy of science and can’t necessarily fault Popper here. But more modern works (such as Gelman and Hennig 2017) discuss objectivist and subjectivist approaches and do not state that the subjective interpretation of probability is both out-moded and unnecessary. Hence we must conclude that they do not wholly accept Popper’s criticisms. We recognize that this is an “appeal to authority” argument which may be flawed, but we don’t feel sufficiently qualified to contribute to the objectivist vs subjectivist debate- we simply observethat the matter does not appear to be closed.

Popper (1984) is mainly concerned with to process of testing theories in physics. Again our use case is not limited to this - it’s about estimating climate sensitivity. Faced with a number of estimates of LGM cooling from different studies, how to we reach a consensus estimate when experts disagree on the merits of different studies? It’s not clear exactly how to do this in an objectivist point of view, especially when other factors (rivalries, personalities, egos, biases) might enter in to it. All we can do here is argue for transparency in decision making.

6. One could argue that we should not worry about words, because “belief” could be interpreted as “estimate” or “information” or “knowledge” or “assumption”. However, I suggest that it is unhelpful to use “belief” in a way that differs radically from its everyday usage. This is epitomised by the biblical story of Doubting Thomas: “Jesus saith unto him, Thomas, because thou hast seen me, thou hast believed: blessed are they that

have not seen, and yet have believed." (John, 20: 29). I suggest that in science, we tend to side with Thomas rather than with Jesus - we tend to demand the evidence and to avoid belief without it.

We are using it in sense of *rational* belief - see Belief, credence, and norms | Philosophical Studies (springer.com) This paper does not discuss religious belief.

We'd also refer to the Stanford Encyclipedia of Philosophy entry on formal belief which discusses both subjective bayesian probability theory and personal questions of faith https://plato.stanford.edu/entries/formal-belief/ Perhaps see also https://plato.stanford.edu/entries/epistemology-bayesian/

7. Another problem of using the term "prior belief" for an element on the right-hand side of the Bayesian equation is that, if we are consistent, the term on the left-hand side of the
equation is then a "posterior belief". However, in this paper and elsewhere, the implication is
that the result of the Bayesian process is an objective result, rather than just a belief – that,
somewhere along the line, a subjective belief is transformed into an objective estimate. Objective theories of probability avoid this problem..

.We've removed instances of "posterior belief". As another reviewer noted, we should not even regard the posterior as an "updated prior", simply the result of the Bayesian process given a set of priors and a model.

8. There is much reference in the paper to "expert judgement" but expert judgement is informed by past experience and its accompanying evidence. Moreover, it is not derived subjectively but through participation in the objective work of the scientific community.

Expert judgment is of course informed by experience and evidence, and it would be ideal if experts derived only via participation in the objective work of the scientific community. As it stands, all experts could have more or less equal access to the published scientific literature- and yet disagreement would persist. It is our goal here to propose methods for the world as it is, not necessarily as it should be.

Perhaps a useful way to think about this is in terms of a hierarchy of models. Why don't we know LGM cooling? We don't know which published estimate to believe. We don't know the proper forward model that converts proxy reconstructions to global mean temperature. There is uncertainty in the proxy measurements. And so on and so on. There is uncertainty at each level. In a perfectly objective world, we'd be able to delve arbitrarily deep into the hierarchy, allowing evidence to determine the posterior distribution of all hyperpriors. However, for tractability, the model must be truncated

somewhere- to paraphrase Newton, we must let ourselves stand on the shoulders of giants/

9. I think the only example of “belief” in this paper is where an “expert” persists in making a
judgement despite evidence to the contrary. I think this is rare – usually there is objective
evidence for a judgement, even though the evidence is incomplete. A good scientist recognises that it is incomplete and is open to new evidence.
More generally, a good scientist holds his/her views tentatively and hypothetically, recalling that scientific progress takes place through the replacement of one false hypothesis by a better (but probably false) hypothesis. Consequently, a good scientist tries not to “believe” anything but to work via a series of hypotheses and assumptions and their testing.

Not all scientists will be familiar with all of the evidence, and some may be over-confident. Part of the motivation here is to find a way of combining the expert judgements of many scientists, relying on the observation that the scientists who are not open to new evidence are a minority group, and so will only have a small impact on the end consensus result.

Detailed comments

11. l.7. Here and many other places. “beliefs”. See General Comments above.
We’ve replaced “beliefs” in most places

12. l.9. Here and many other places. “subjective”. See General Comments above.
For the reasons above, we’ve kept “subjective”. We must decide what evidence to use, assess its quality, choose a model (or candidate models) to interpret it, specify priors on model parameters, and decide how different lines of evidence relate to one another. All of these are decisions that must be made, and therefore we feel “subjective” is appropriate.

13. l.19, eq.(1). What is M 0 ? - the climate system, a model of the climate system, or the
simple energy balance model? If the last, then is the RHS of (1), i.e. including ΔN, different?

We’ve more clearly defined M0

14. l.33: “aerosols”. Net cooling in response to aerosols?

Reworded to “We also have the evidence of the planet itself, which has been steadily warming in response to net anthropogenic forcing, which includes not just emissions of $\rm{CO}_2$ but of other greenhouse gases and aerosols as well.”

15. l.49 and l.170: “knowledge”. See General Comments above.
We’ve removed most references to “belief”

16. l.61-622, eq.(3) and following line. If P(Θ) is a belief, then P(Y|Θ) must also be a belief
(a posterior belief). See General Comment 7 above.
This section has been rewritten

17. Fig.1(a). Axes need labels.

Fixed.

18. l..120, equation. This is not very clear. C is not defined.
This section has been rewritten to emphasize the “probability mass” concept- please see response to R1 above.

19.
l.161-162. Sentence “These incorporate expert judgement ...”. These are normally objective, not subjective, i.e. they are inter-subjectively shared and criticised. This is fundamental in science.

Yes, expert judgment should be shared and criticized (and this is one of the goals of this paper) but this does not necessarily make such judgements, especially about an uncertain quantity, “objective”- merely that frameworks such as this one that seek to interrogate and synthesize these judgments using evidence are necessary.

20.
l.166-167: “well-informed scientist”. Again, informed by objective information.21.

Unfortunately, well-informed scientists may still have imperfect knowledge or different opinions on the literature

l.217-218: “Why do these two distributions not overlap substantially?” They appear to overlap substantially - they are well within each other’s one-sigma points.
This is fair, and we’ve removed this. We’re only trying to illustrate that the two distributions overlap more if a model with state dependence in the LGM is used.
.

22.
l.229: “odds”. This is another word associated with a subjective theory of probability,

and best avoided if you adopt an objective approach.

This is likely another example of where we disagree due to hierarchy truncation. Yes, an expert should allow evidence to inform her/his judgment of model odds. But that would require "going another step down" in the hierarchy and having a scientific debate over the prior odds, which would in turn require a debate over the evidence informing those prior odds, and so on and so on. We feel it's better, in a tractable analysis, to simply clearly specify these priors.

23.
l.237: "definite". What does definite mean here? Does it mean "certain"? If so, this would not be a scientific statement - uncertainty is all-pervasive in science. If you remove
"definite" from this sentence, do you not conclude that state dependence is likely?

We've rewritten this section and it no longer appears.

24.
l.242: "We are not arguing that this is the objectively "correct" way to combine the Last Glacial Maximum reconstructions with historical observations." Given my comment above, it is not clear what you are arguing here.

We've rewritten this section and it no longer appears.

25.
l.250: "relying on a community of experts". Yes! - this makes it objective - this is how we do science - inter-subjectively shareable and criticisable.

This is how science *should* be done, but it is not how it *is* done- at least on timescales necessary for publishing assessments! Experts are often unable to reach consensus decisions, and thus a framework that incorporates potentially subjective prior information is necessary. We feel that making decisions transparent is the first step toward such important criticism.

26.
l.263: "assume". Yes - so these are hypotheses (to be tested), not beliefs.
Yes, as we point out in the previous section, the Bayesian model evidence allows us to assess models in light of the evidence. Our ability to reject or accept a hypothesis, however, depends strongly on the prior odds, and is - in a truncated model hierarchy- subjective.

27.
l.286: "prior assumptions: Yes - much better! You can assume something without believing it.
Ok, we've kept this

28.
l.290: “accurate”. Meaning exact? Unusually, accuracy means a quantification of the Uncertainty.
Replaced with “Similarly, we might set the prior on μ using the result of a single published study (say, for example, ΔT from T20). “

29.
l.294: “belief”. At no point in the discussion contained in this paragraph do you need to “believe” anything - you are making certain assumptions or posing certain hypotheses, and then
testing their consequences.

We’ve removed “belief”

30.
l.337: “the prior beliefs of two hypothetical experts”. Or you could say just two hypotheses?
31.
l.341: “However, some experts may not be so open-minded ...”. So, are you are saying that there are closed-minded experts who “believe” things and open-minded experts who make hypotheses?

Perhaps we’re saying that there are some experts who might be considered by some to be unscientific. But we need to include their views to give a consensus estimate. Also it’s not a clear-cut thing - there’s a sliding scale depending on how narrow their priors are.

32.
l.343-345: “Expert C’s confidence remains unshaken ...” and following sentence.
This is fundamental to how science works. You are saying that Expert C is not influenced by
evidence and so is not behaving rationally/scientifically. In (good) science, we suspend belief and act tentatively and hypothetically.

Yes, the idea is that this method allows you to include the views of all experts, whether they are willing to modify their views or not. This is what IPCC does, and we’re trying to do the same, but in a more transparent way. Hopefully their contributions will be downweighted but not entirely ignored. This avoids you having to screen out people who you think might be overconfident, something that can be problematic, when building a community assessment.

33.
l.348-349: “The narrowness of C’s prior ...”. It’s OK to have a narrow prior, if all the evidence you have (at present) points in that direction, but it is prudent to assume that there is some possibility (low probability) of a gross error, because of some effect that

Towards robust community assessments of the Earth's climate sensitivity

Kate Marvel[1] and Mark Webb[2]

[1]NASA Goddard Institute for Space Studies, New York, NY, USA
[2]Met Office Hadley Centre, Exeter, UK

Abstract. The eventual planetary warming in response to elevated atmospheric carbon dioxide concentrations is not precisely known. ~~This climate sensitivity~~ The uncertainty in S ~~depends primarily on the~~ primarily results from uncertainties in net physical climate ~~feedbacks~~feedback, usually denoted as λ. Multiple lines of evidence can constrain this feedback parameter: proxy-based and model evidence from past equilibrium climates, process-based understanding of the physics underlying changes, and recent observations of temperature change, top-of-atmosphere energy imbalance, and ocean heat content. However, despite recent advances in combining these lines of evidence, the estimated range of S remains large. Here, using a Bayesian framework, we discuss three sources of uncertainty: uncertainty in the evidence, structural uncertainty in the model used to interpret that evidence, and differing prior knowledge and/or beliefs, and show how these affect the conclusions we may draw from a single line of evidence. We then propose ~~a method~~ strategies to combine multiple ~~estimates of the evidence, multiple multiple explanatory models, and the subjective assessments of different experts in order to arrive at an assessment of λ (and hence, climate sensitivity S) that may be rapidly updated as new information arrives and truly reflects the existing community of experts~~lines of evidence. We end with three recommendations. First, we suggest a Bayesian random effects meta-analysis be used to estimate the evidence and its uncertainty from published literature. Second, we advocate that the organizers of future assessments clearly specify an interpretive model or group of candidate models,in the latter case using Bayesian model averaging to more heavily weight models that best fit the evidence. Third, we recommend that expert judgment be incorporated via solicitations of priors on model parameters.

1 Introduction

When a radiative forcing ΔF is applied to the climate system, it induces a radiative imbalance ΔN at the top of the atmosphere and a response ΔR of the system itself. To first order, $\Delta R = \lambda \Delta T$, where ΔT is the change in global mean surface temperature. The feedback parameter λ thus measures the additional radiative flux density exported to space per unit warming. On sufficiently long timescales the climate comes into equilibrium ($\Delta N = 0$), internal variability is negligible and we can write a simple energy balance model (denoted M_0) for the climate system:

$$\mathcal{M}M_0 : \Delta N = \Delta F + \lambda \Delta T. \quad (1)$$

In the special case where the radiative forcing results from a doubling of atmospheric CO_2 relative to its preindustrial concentration of 280ppm ~~and the system is allowed to come into equilibrium ($\Delta N = 0$), then the internal variability term is negligible and the~~ ($\Delta F = F_{2\times CO_2}$) the resulting temperature change defines the ~~equilbrium~~ equilibrium climate sensitivity S:

$$\Delta T_{2\times CO2} \equiv S \equiv -\frac{F_{2\times CO2}}{\lambda}. \quad (2)$$

~~S~~ S is often used as a metric to quantify expected warming in response to radiative forcing, but has remained stubbornly uncertain even as climate models have improved and become more sophisticated. A 2020 community assessment (~~(Sherwood et al., 2020)~~ Sherwood et al. (2020), hereafter S20) reduced this range using multiple lines of evidence, but the recent IPCC report ~~(Forster, 2021)~~ Forster (2021) assessed only "medium confidence" in the upper bound. ~~It is therefore imperative to reduce the uncertainty and enhance confidence in a quantity so crucial to climate science and policy.~~ Is it possible to further narrow the estimated range of S, and can we increase our confidence in this result?

~~S~~ S is determined by the net feedbacks λ at equilibrium and in response to doubled CO_2. While these are unobservable in the current system, in which CO_2 has not yet doubled and which is out of equilibrium, there exist several lines of evidence that might constrain λ. We have some process-based understanding of individual feedback processes and their correlations derived from observations and basic physics. We also have the evidence of the planet itself, which has been steadily warming in response to ~~anthropogenic~~ net anthropogenic forcing, which includes not just emissions of CO_2 ~~,~~ but of other greenhouse gases ~~, and aerosols~~ and aerosols as well. Finally, we have proxies that provide evidence about equilibrium climates of the past. S20 attempted to synthesize these three lines of evidence, ~~incorporating the judgement of many experts, and arrived~~ arriving at constraints on climate sensitivity that narrowed the former range.

In S20, the spread in S arose from reported and assessed uncertainty in historical observations and paleoclimate reconstructions, expert judgement about the uncertainty of physical processes, and the use of different priors on λ and/or S. IPCC AR6 assessed confidence in the range of S based on support from individual lines of evidence, and the medium confidence assessed was in large part due to the fact that not all lines of evidence supported the same upper bound. By contrast, S20 sought to provide a robust estimate by combining lines of evidence in a coherent Bayesian framework. However, S20 used baseline priors and estimates of the evidence and investigated the impact of alternate choices as sensitivity tests rather than attempt to combine multiple priors, estimates, and expert judgements into a single posterior probability distribution. In both IPCC AR6 and S20, as in almost all previous assessments, the means by which disagreements among experts were resolved or handled was not necessarily made transparent. This paper presents some lessons learned by two authors of S20 and attempts to chart a way forward. ~~Our goals are to 1) better understand the sources of uncertainty 2) understand~~

Our goal is to understand where unavoidable subjective decisions enter in to the analysis and ~~3)~~ to present a framework for systematically and fairly incorporating the subjective judgements of multiple experts. Ultimately, we seek to create a framework in which expert judgement is incorporated in the form of clearly specified priors.

The paper is organized as follows. In section 2, we review the basic Bayesian analysis framework. Sections 3, 4, and 5 discuss evidence, structural, and prior uncertainty, respectively. In these sections, we use a single line of evidence– paleoclimate data from the Last Glacial Maximum– to illustrate how these sources of uncertainty shape estimates of climate feedbacks and

sensitivity. In Section 6 we show how these sources of uncertainty affect constraints derived from multiple lines of evidence. In Section 7 we propose a new method for combining multiple published studies and multiple models, which may be used in the future to arrive at a robust community assessment of climate sensitivity. Finally, in section **??** we discuss possible generalizations and extensions.

2 Analysis framework

~~As in S20, we use a Bayesian framework in which researchers express their beliefs about model parameters in terms of probability distributions. Bayesian statistics is both praised and criticized for its inherent subjectivity (see, e.g. (Gelman et al., 1995)). The framework requires researchers to specify prior distributions, which in general reflect some degree of knowledge (Gelman et al., 2017). Different researchers may quite reasonably have different beliefs if they do not have access to the same observations or disagree with one another about the credibility of lines of evidence. The requirement to specify their priors forces researchers to be explicit about assumptions and pre-existing beliefs that might otherwise be implicit but ignored. After all, *every* statistical analysis contains subjective choices to some extent, whether is is the choice of the p-value threshold in frequentist statistics or the selection of priors in Bayesian frameworks. Moreover, every statistical analysis depends on the model used to interpret the evidence. Here, we show where unavoidable choices enter into the analysis framework, differentiate between uncertainty in the model, uncertainty in the evidence, and uncertainty in the parameters, and suggest strategies to handle and reduce these uncertainties in future work. We will introduce Bayesian hierarchical modeling approaches to weight alternative models, evidence sources, and expert beliefs. Our hope is to lay the groundwork for future syntheses of evidence and to constrain the use of "expert judgement" within a clearly defined, well-constrained framework.~~ Bayes' Theorem can be written as

~~Bayes' Theorem states that~~

~~$$P(\Theta|Y) = \frac{P(Y|\Theta)P(\Theta)}{P(Y)}.$$~~

$$P(\Theta|Y,M) = \frac{P(Y|\Theta,M))P(\Theta|M)}{P(Y|M)}. \quad (3)$$

~~Here, we will define these terms as they apply to the problem of estimat The term $P(Y|\Theta)$ is the likelihood, defined as the probability of the data given some putative values of the parameters Θ. The data~~ Here, we will define these terms as they apply to the problem of estimating climate sensitivity.

Evidence The evidence Y ~~consist~~ used to constrain climate sensitivity consists of the global mean temperature change ΔT in response to a forcing ΔF as well as, in non-equilibrium states, the net energy inbalance ΔN. We have estimates of these quantities for the historical period (derived from observations and models) and for past climate states (derived from paleoclimate proxies and models), and Y therefore consists of multiple lines ~~of evidence~~ $Y_1 \ldots Y_n$~~;~~. For example,

S20 used ~~direct observations, models of varying complexity,~~ process-based understanding of underlying physics, recent observations, and proxy-based reconstructions of past climates ~~. In section 3 we will discuss how differing interpretations of the evidence introduce uncertainties into the analysis.~~

~~The~~ to assess S.

Model The model M codifies how we interpret the evidence Y. It specifies the parameters Θ ~~are specified by an underlying generative model $\mathcal{M}$ for the data~~whose posterior distributions we estimate. For example, ~~the simplest~~ in the simple energy balance model ~~$\mathcal{M}_0$ (Equation 1) contains a single parameter (~~denoted M_0, there is only one parameter and $\Theta = \lambda$~~).~~ ~~This model carries with it the implicit assumptions that the feedback parameter is constant in time, that a unit of global mean radiative forcing always produces the same temperature change, and that the radiative imbalance ΔN always has unit efficacy. All of these assumptions have been challenged in the literature (e.g.(Andrews et al., 2018; Winton et al., 2010; Forster, 2) . Other, more complex models with additional parameters are possible, and we will discuss these in Section 4.~~

~~Finally, the term $P(\Theta)$ corresponds to subjective prior beliefs about the values~~ The model determines the **likelihood** $P(Y|\Theta, M)$ of observing the data given particular values of the parameters Θ. We discuss methods for calculating this likelihood in Section 3.1.

Prior The **prior** probability distribution $P(\Theta|M)$ reflects prior beliefs or knowledge about the model parameters Θ. For example, in the simple model M_0, the community assessment S20 adopted a uniform prior on λ as a baseline choice, choosing not to rule out net positive feedbacks (and therefore an unstable climate) *a priori*. Both the geological evidence and process understanding presented in Section 3 of S20 effectively rule out both positive and extremely negative feedbacks, and thus an alternate prior reflecting this physical knowledge might be a normal distribution $N(\mu, \sigma)$ with mean μ =-1.30 and standard deviation $\sigma = 0.44$.

This framework allows us to use our prior understanding of the parameter values to calculate the **posterior** probabilities $P(\Theta|Y, M)$ of the model parameters given the evidence. This posterior can be updated as new evidence becomes available.

Bayesian statistics is both praised and criticized for its inherent subjectivity (see, e.g. Gelman et al. (1995)). But *all* statistical analyses depend on prior knowledge and interpretive models, whether implicit or explicit. The Bayesian framework merely makes clear where unavoidable subjective decisions enter the analysis. ~~While in theory sufficient evidence should update the priors of all reasonable analysts and result in similar posterior estimates, in practice sparse evidence and strongly held beliefs mean consensus may not be reached after the analysisis performed. In Section 5 we will discuss how these subjective prior beliefs affect our estimates of~~

Figure 1 summarizes the decisions that must be made in any Bayesian analysis of climate feedbacks. First, the analyst must decide what constitutes "evidence". This requires an assessment of the ~~feedback parameter λ.~~ literature assessing ΔT, ΔF, and ΔN for each line of evidence. Second, the analyst must specify a model (and its parameters Θ) in order to interpret that evidence. For example, the model M_0 assumes the feedback parameter is time- and state-independent, and thus estimating it from the past is a reliable guide to the hypothetical future under doubled CO_2. Finally, the analyst must clearly specify her or his priors on the model parameters.

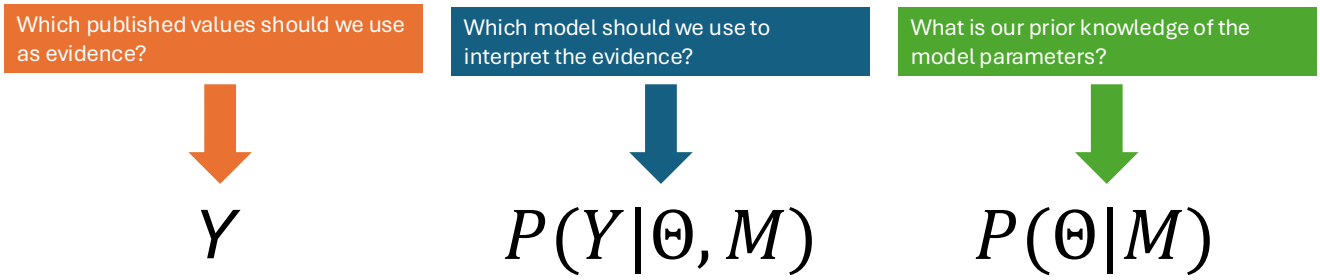

Figure 1. Schematic of unavoidable subjective decisions in an analysis of climate feedbacks.

3 ~~Evidence uncertainty~~

~~The first type of uncertainty we highlight is uncertainty in the *evidence*. Evidence-based constraints on climate sensitivity (S) or feedback parameters (λ) are generally derived from estimates of temperature response, energy imbalance, and radiative forcing. These may come directly from observations (e.g. the instrumental warming record or measured ocean heat uptake), from model calculations which are informed by observations and theory (e.g. radiative forcing) or from observational evidencewhich is interpreted using modeling assumptions~~In the following sections, we show how different reasonable choices about evidence, models, and priors can lead to very different posterior distributions for λ (and hence climate sensitivity S) given a single line of evidence.

~~In S20, the~~

3 Evidence uncertainty

The strongest constraints on ~~the upper bound of warming~~ equilibrium climate sensitivity in S20 were derived from paleoclimate evidence~~. In this section, we will show how uncertainty in the evidence affects our confidence in those constraints. The~~ , and the closest equilibrium climate to ~~ours~~ the present is the Last Glacial Maximum (LGM) approximately 21,000 years ago. Reconstructions ~~(Annan and Hargreaves, 2013; Bereiter et al., 2018; Friedrich et al., 2016; Holden et al., 2009; von Deimling et al., 2006; Shaku~~ Annan and Hargreaves (2013); Bereiter et al. (2018); Friedrich et al. (2016); Holden et al. (2009); von Deimling et al. (2006); Shakun et a or model-based estimates ~~(Braconnot et al., 2012; Kageyama et al., 2021)~~ Braconnot et al. (2012); Kageyama et al. (2021) of the global mean temperature change ΔT and the radiative forcing ΔF have been used to calculate the global mean feedbacks λ inferred from this period. Neither of these "observed" quantities is precisely known. For example, multiple, seemingly incompatible, estimates of the LGM global mean cooling ΔT are available in the published literature ~~(Annan and Hargreaves, 2013; Holden et al.,~~ Annan and Hargreaves (2013); Holden et al. (2009); Shakun et al. (2012); von Deimling et al. (2006); Friedrich et al. (2016); Hansen et al . These are derived from climate models participating in the Paleoclimate Model Intercomparison Project (PMIP ~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)) and combinations of models and various proxies, and are often in conflict with one another.

We will illustrate the impact of this uncertainty by comparing the evidence used in two recent studies. S20 used expert judgement applied to a literature review to estimate $\Delta T = -5$K with a 95% confidence interval of (-3.0K, -7.0K). However, a contemporaneous study using ~~an updated~~ a new temperature reconstruction (Tierney et al 2020 ~~(Tierney et al., 2020)~~ Tierney et al. (2020), hereafter T20) estimated both colder (mean -6.1K) and less uncertain (with a 95 % highest posterior density interval of -6.5 to -5.7 K) values for LGM cooling. We note that the two studies are not exactly comparable: S20 represents a community assessment of evidence that took into account a broad range of evidence and uncertainties, whereas T20 was a single study. The temperature estimates in T20 may also be cold-biased and overconfident due to reliance on a prior derived from a single climate model Annan et al. (2022). However, in order to illustrate evidence uncertainty, we here treat S20 and T20 as different reasonable estimates of ΔT and ΔF over the LGM. We discuss methods for incorporating estimates such as T20 in expert assessments in Section 7.1.

The two studies S20 and T20 also differ in their estimates of the radiative forcing that led to this temperature change. Both agree that it was colder 21,000 years ago because a change in orbital forcing, while negligible in the global mean, led to the development of large, reflective ice sheets in the northern hemisphere and lower levels of atmospheric greenhouse gases. The forcings associated with orbital changes ~~(Kageyama et al., 2021)~~ Kageyama et al. (2021) and CO_2 ~~(Siegenthaler et al., 2005)~~ Siegenthaler et al. (2005) are relatively well-constrained; the forcings from other well-mixed greenhouse gases ~~(Loulergue et al., 2008)~~ Loulergue et al. (2008) and ice sheets less so but still informed by proxy and model evidence (Section **??**), and those from dust ~~(Mahowald et al., 2006; Albani and Mahowald, 2019), ,~~ Mahowald et al. (2006); Albani and Mahowald (2019), other aerosols, and vegetation ~~(Köhler et al., 2010)~~ Köhler et al. (2010) highly uncertain. While S20 estimated total radiative forcing in the LGM to be N(-8.43, 2) W m^{-2}, T20 use a best estimate of -6.8 W m^{-2} with a 95% confidence interval of ~~9.6~~ -9.6 to -5.2 W m^{-2}.

~~Some uncertainty in the climate feedback λ or the climate sensitivity S therefore derives from uncertainty in the evidence used to constrain those parameters. Here, we will define *evidence* uncertainty as uncertainty in the joint probability density of the evidence Y. This is defined as $Y_{LGM} = (\Delta T, \Delta F)$: the global cooling and radiative forcing during the LGM. Assuming the uncertainty in ΔT is independent of the uncertainty in ΔF, the resulting joint probability density functions derived from~~ Contour lines in Figure 2a show the joint probability distribution (assuming uncorrelated errors) $\rho(\Delta T, \Delta F)$ as reported by S20 (black) and T20 ~~are shown in Figure 2(a) . The estimates used in S20 are Gaussian and have large uncertainties (blackcontours) , while uncertainties in both temperature and radiative forcing are smaller in T20~~ (red~~contours). T20 also treats non-greenhouse gas forcing as non-Gaussian~~).

~~Panel (a): joint evidence distributions for ΔT and ΔF used in Sherwood et al (black contours) and Tierney et al (red contours). Structural uncertainty is illustrated using solid lines (corresponding to fixed values of λ using the model $\mathcal{M}_0$) and dashed lines lines (corresponding to fixed values of λ and α using the model $\mathcal{M}_\alpha$). (b): Likelihoods as a function of λ and given S20 (black lines) or T20 (red lines) evidence and different values of the state dependence α. (b): Resulting likelihoods for λ given the evidence from S20 (black) and T20 (red) and different values of the state dependence parameter α. Likelihoods derived using the simple energy balance model ($\alpha = 0$) are highlighted by thick lines.~~

~~Using the simple energy balance model (Eq. 1) and setting $\Delta N = 0$, the forcing is proportional to the temperature change , with the only parameter the net feedback λ. This means that if we knew the temperature change~~ Rather than exact measurements of the temperature change and radiative forcing~~*exactly*, we would know the feedback λ, and if we knew the forcing at doubled CO_2, we would then know the climate sensitivity exactly. However, we do not know the exact cooling and radiative forcing during the LGM, but rather a~~ , our evidence Y consists of estimates of the joint probability density ~~$\mathcal{J}(Y)$ for both. The model $\mathcal{M}_0$ imposes the requirement that given a fixed~~ $\rho(\Delta T, \Delta F)$.

3.1 Calculating the likelihood

The likelihood of "observing" this probability density for any given value of the feedback parameter λ is determined by the model, which dictates the relationship between λ, ~~all pairs of $(\Delta T, \Delta F)$ lie on the~~ ΔT and ΔF. For example, the simple energy balance model M_0 constrains all possible pairs of $(\Delta T, \Delta F)$ to line on a line with slope ~~λ. Integrating the joint probability density along that line results in~~ $-\lambda$. Intuitively, the value of $-\lambda$ that maximizes the likelihood is the slope of the line that passes through through the greatest probability density. These maximum likelihood estimates are shown as straight lines in Figure 2a.

We therefore define the likelihood of ~~the evidence given~~ $\rho(\Delta T, \Delta F)$ for any λ ~~(Figure 2(b)):~~ as the probability mass along the curve C described by the energy balance model with fixed λ:

$$P(\Delta T, \Delta F Y|\lambda) \propto \int_C \mathcal{J}\rho(\Delta T, \Delta F) ds$$
$$C : 0 = \Delta F + \lambda \Delta T$$

If the joint evidence is ~~Gaussian~~a multivariate normal distribution (as it is in S20), this leads to an exact analytic expression for $P(Y|\lambda)$ (Appendix 1). Otherwise, the integral can be computed numerically. The resulting likelihood functions are shown as thick lines in Figure 2b.

3.2 Climate sensitivity estimates depend on the evidence

Clearly, the constraints placed on the climate feedback by the Last Glacial Maximum depend on our estimates of the temperature difference and the radiative forcing that caused it. Using S20 evidence, this energy balance model, and a uniform prior $P(\lambda) = U(-10, 10)$, we find that the most likely value of the feedback parameter is $\lambda = -1.7$ $\mathrm{Wm^{-2}K^{-1}}$ (thick black line, Figure 2~~(b)~~ b) with a 5-95% range of (-3.37, -1.09) $\mathrm{Wm^{-2}K^{-1}}$. ~~This corresponds to a 5-95% range of (1.17K, 3.69K) for the climate sensitivity S (assuming, as in S20, that $F_{2\times CO_2} \sim N(4.0, 0.3)$.~~ Using T20 evidence, the most likely value is $\lambda = -1.1\,\mathrm{Wm^{-2}K^{-1}}$ (~~(~~thick red line, Figure 2~~(b)~~b). The 5-95% range is (-1.49, -0.87) $\mathrm{Wm^{-2}K^{-1}}$~~for λ and (2.61, 4.72) K for S. Clearly, the constraints placed on the climate feedback by the Last Glacial Maximum depend on our estimates of the temperature difference and the radiative forcing that caused it.~~ .

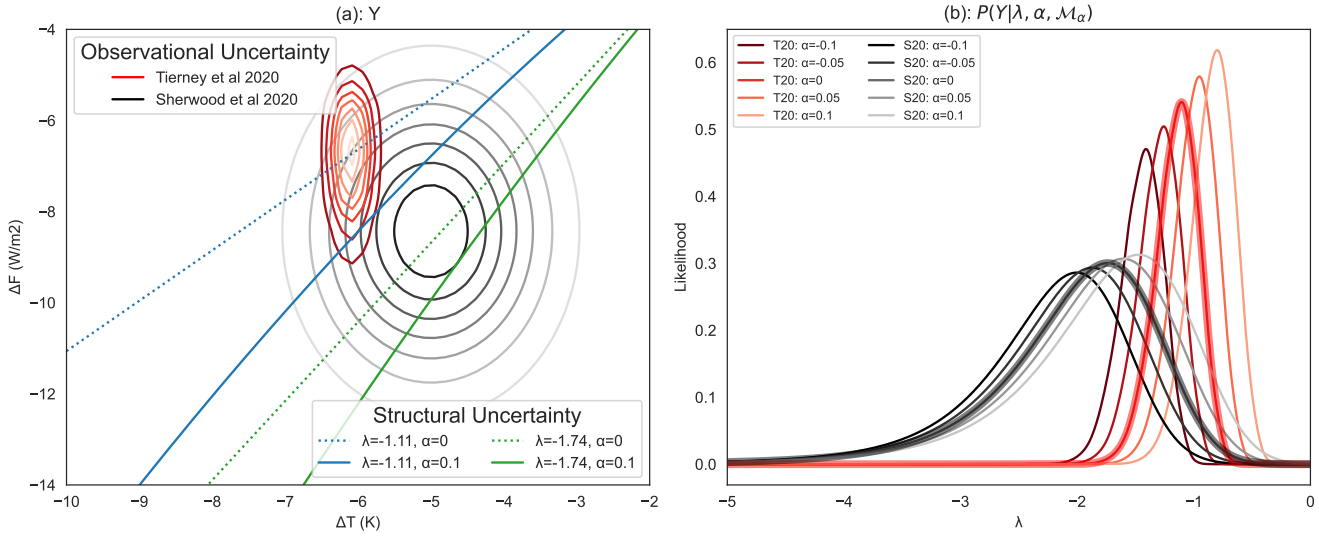

Figure 2. Panel a: joint evidence distributions for ΔT and ΔF used in Sherwood et al (black contours) and Tierney et al (red contours). Structural uncertainty is illustrated using solid lines (corresponding to fixed values of λ using the model M_0) and dashed lines lines (corresponding to fixed values of λ and α using the model M_α). b: Likelihoods as a function of λ and given S20 (black lines) or T20 (red lines) evidence and different values of the state dependence α. b: Resulting likelihoods for λ given the evidence from S20 (black) and T20 (red) and different values of the state dependence parameter α. Likelihoods derived using the simple energy balance model ($\alpha = 0$) are highlighted by thick lines.

For simplicity, here we calculate the likelihood ~~$P(\Delta T, \Delta F|\lambda)$~~ $P(Y|\lambda)$, and use the resulting posterior ~~$P(\lambda|\Delta T, \Delta F) \propto P(\Delta T, \Delta F|\lambda)P$~~ $P(\lambda|Y) \propto P(Y|\lambda)P(\lambda)$ to calculate S (Appendix 2). This neglects the small correlation between ΔF and the forcing at doubled CO_2, but this simplification does not substantially affect our results (Appendix 3).

Using S20 evidence from the LGM, we find a 5-95% range of (1.17K, 3.69K) for the climate sensitivity S (assuming, as in S20, that $F_{2\times CO_2} \sim N(4.0, 0.3)$). Using T20 evidence, the 5-95% range for S is (2.61K, 4.72K).

4 Structural uncertainty

~~Many~~ Thus far, we have relied on the simple energy balance model to interpret the LGM evidence. However, many recent studies (e.g.~~(Rohling et al., 2018; Stap et al., 2019; Friedrich et al., 2016)~~Rohling et al. (2018); Stap et al. (2019); Friedrich et al. (2016); Renc) suggest that ~~the simple model $\mathcal{M}_0$~~ M_0 might not be appropriate for past climates due to the dependence of the feedbacks on the background climate state. If the relationship between temperature change and radiative forcing is nonlinear, then the feedbacks in a past cold climate should not be treated as identical to those in a future warm one. To model this background temperature dependence, we might use an alternate model that includes a second-order term in the radiative response

$$\mathcal{M}_{\alpha} : 0 = \Delta F + \lambda \Delta T + \frac{\alpha}{2} \Delta T^2 \quad (4)$$

where $\alpha = \partial\lambda/\partial(\Delta T)$ is an additional parameter reflecting the background state dependence ~~(Sellers, 1969; Caballero and Huber, 2013;~~ Sellers (1969); Caballero and Huber (2013); Budyko (1969); Sherwood et al. (2020). Intuitively, nonzero values of α change the relationship between the paleoclimate evidence and the feedback parameter λ. This, in turn, makes the evidence more or less likely given a value of λ. For example, if $\alpha = +0.1$ (which translates to a change in feedback of -0.5 $\mathrm{Wm^{-2}K^{-1}}$ at a cooling of -5 K), the most likely value of λ is not the same as the most likely value of λ assuming $\alpha = 0$ (dotted and solid lines, Figure 2a). In this case, the likelihoods (Figure 2b) are calculated by integrating the joint probability distribution for ΔT and ΔF along the curve defined by Eq. 4, and depend on the value of the state dependence parameter α.

If α is not a fixed value but an unknown parameter, then the evidence can constrain only the joint distribution of $\Theta = (\lambda, \alpha)$. Obviously, in order for the climate of the past to tell us anything about the climate of the future, we must have some information about how they relate to one another.

There is no limit to the complexity of models we might use to interpret the evidence of the LGM. We might allow for both non-unit forcing efficacy and state dependence~~, for example, or~~ . We might assign different efficacies to different forcing agents, or allow the parameter α to bifurcate at lower temperatures. We might also include an additive pattern effect $\Delta\lambda$ that reflects differences in the spatial pattern of temperature change in the LGM and the pattern of warming expected at elevated CO_2 concentrations (e.g. Cooper et al. (2024)).

~~The interpretive model~~ Regardless of the interpretive model used, it is both required for analysis and subjectively chosen by the analyst. Different reasonable analysts might make different choices about the model to use. This means that the choice of model is an important source of uncertainty that must be clearly specified or quantified. There is, however, one more source of uncertainty to discuss. Even given a single model, for example ~~$\mathcal{M}_{\alpha}$~~M_{α}, our degree of confidence in the constraints placed by paleoclimate evidence on the feedback parameter λ necessarily reflects our prior knowledge of the state dependence of climate feedbacks. It is to this prior uncertainty that we turn in Section 5.

5 Prior uncertainty

Once a model is specified, we would like to use the evidence to tell us something about its parameters Θ. Bayes' theorem says that the posterior ~~distributions- essentially, our degree of belief in the values of those parameters-~~ distributions of the parameters are simply obtained by multiplying the likelihood by ~~priors:~~ prior probability distributions reflecting our pre-existing beliefs ~~.~~ ~~These~~ and/or knowledge. These priors incorporate expert judgement, the results of other analyses, and knowledge of physical processes. Posterior distributions of individual parameters can depend strongly on prior knowledge of all parameters. For example, Figure 3~~(a)~~ a shows the joint posteriors for the feedbacks λ and the state dependence α assuming the model ~~$\mathcal{M}_{\alpha}$~~M_{α}, the temperature and radiative forcing values reported in S20, and uniform priors on both parameters. In the absence of any physical knowledge about these parameters, the joint posterior is not very informative. In fact, considerable posterior weight

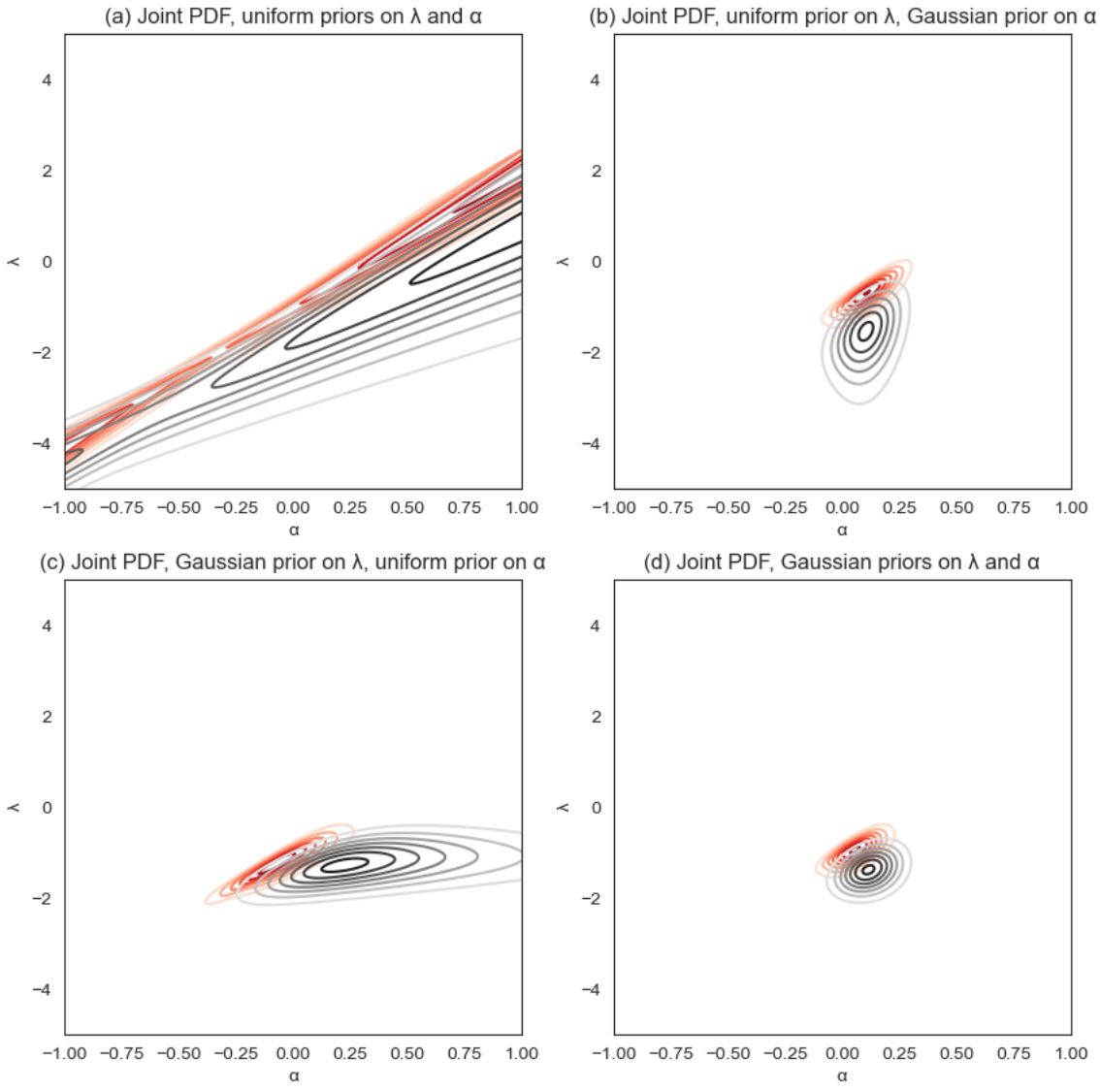

Figure 3. Joint posteriors for the feedback parameter λ and the state dependence α under different priors: ~~(~~a ~~)~~Uniform priors on both parameters ~~(~~b ~~)~~Uniform prior on λ, Gaussian prior from expert judgement of published literature (used in S20) on α (c) Gaussian prior from process evidence (used in S20) on λ, uniform prior on α (d) Gaussian priors (from S20) on both.

is placed on extremely large positive values of α and positive λ~~:~~, which would make negative climate sensitivity appear more likely than most scientists would consider credible. A well-informed scientist, however, is unlikely to think that $\alpha = 1$ (which implies an enormous mean change in feedback of -5 W m^{-2} K^{-1} for 5K of glacial cooling) is just as likely as $\alpha = 0$ (implying no change in feedback). In S20, a prior of ~~$\mathcal{N}(+0.1, 0.1)$~~ $N(+0.1, 0.1)$ was assigned to the state dependence α, reflecting the current state of the literature. This prior ~~knowledge~~ substantially constrains the resulting joint posterior distribution (Figure 3 b). Conversely, imposing a more informative prior on the ~~feedbacks~~ feedback parameter λ, for example by using the process constraints in S20 that result in ~~$\lambda \sim \mathcal{N}(-1.30, 0.44)$~~ $\lambda \sim N(-1.30, 0.44)$, also constrains the joint distribution: positive values of α (i.e., which imply a lower sensitivity in the LGM than at doubled CO_2) receive more posterior weight. Combining the informative priors on both λ and α further constrains the joint posterior (Figure 3~~(d)). Schematic of unavoidable subjective decisions in an analysis of climate feedbacks.~~ d).

~~Figure 1 summarizes the subjective decisions that must be made in any Bayesian analysis of climate feedbacks. First, the analyst must decide what constitutes "evidence". In the LGM example we have presented, this requires subjective assessment of the literature assesing the cooling ΔT and the radiative forcing ΔF. Different highly credible published studies lead to very different likelihoods for λ. Second, the analyst must specify a model (and its parameters Θ) to interpret that evidence. For example, the model $\mathcal{M}_0$ assumes the feedback parameter is time- and state-independent, and thus estimating it from the past is a reliable guide to the hypothetical future under doubled CO2. The model $\mathcal{M}_\alpha$, by contrast, assumes the feedbacks depend on the background climate and introduces a second parameter to reflect state dependence. Finally, the analyst must specify her or his prior beliefs about the model parameters.~~

6 Combining multiple lines of evidence

The examples we have presented thus far have all used ~~paleo evidence from~~ a single line of evidence– paleoclimate reconstructions of the Last Glacial ~~Maximum~~ Maximum– to constrain λ. However, it is not necessary to look back over twenty thousand years to gauge the planet's response to external influences. More recently, a large increase in radiative forcing has resulted in significant global warming and a large ~~imbalance reflected in an increased rate of ocean heat uptake~~radiative imbalance at the top of the atmosphere. To constrain λ with transient historical observations, ~~the~~ we use evidence $Y = (\Delta T, \Delta F, \Delta N)$, where ΔN is estimated from observed changes in ocean heat uptake and/or satellite observations constrained by ocean heat content ~~(Forster, 2016).~~ Forster (2016).

6.1 Historical likelihood

In this three-dimensional joint probability space, the simplest energy balance model ~~$\mathcal{M}_0$~~ M_0 defines a plane rather than a line in evidence space (Figure 4), and the likelihood of the evidence given λ is proportional to the integral over this surface. Figure

(a): Line integral (b): Likelihoods

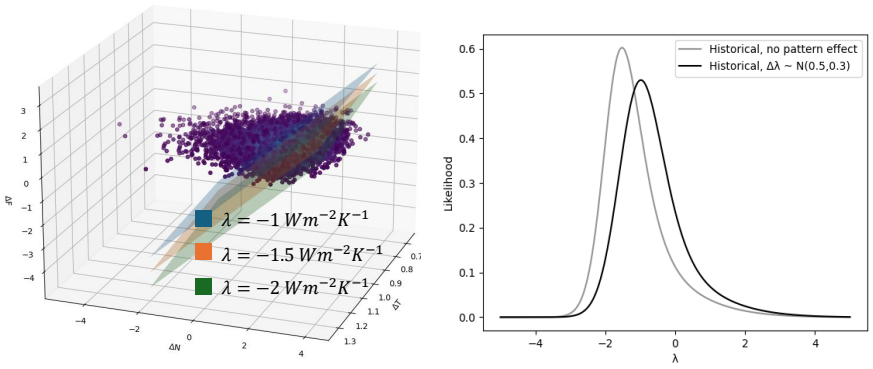

Figure 4. a: Calculating the likelihood of observing the historical evidence used in S20 for a putative value of λ. Each value of λ defines a plane; shown are ~~$\lambda = -1\mathrm{Wm}^{-2}\mathrm{K}$~~ $\lambda = -1\mathrm{Wm}^{-2}\mathrm{K}^{-1}$ (blue), ~~$\lambda = -1.5\mathrm{Wm}^{-2}\mathrm{K}$~~ $\lambda = -1.5\mathrm{Wm}^{-2}\mathrm{K}^{-1}$ (orange) and ~~$\lambda = -2\mathrm{Wm}^{-2}\mathrm{K}$~~ $\lambda = -2\mathrm{Wm}^{-2}\mathrm{K}^{-1}$ (green). The likelihood is the surface integral of the joint PDF along the plane. b: Likelihood for the feedback parameter λ given the simple energy balance model with no pattern effect (gray line) and marginal likelihood for λ given an additive pattern effect with prior $\Delta\lambda \sim N(0.5, 0.3)$.

4 shows the historical evidence reported in S20, in which

$$\Delta T \sim \mathcal{N}(1.03, 0.085) N(1.03, 0.085) \quad (5)$$

$$\Delta N \sim \mathcal{N}(0.6, 0.18) N(0.6, 0.18) \quad (6)$$

and ΔF is calculated using unconstrained aerosol ERFs from ~~(Bellouin et al., 2020)~~ Bellouin et al. (2020) with median 1.83 W m^{-2} and 5-95% range (-0.03, 2.71) W m^{-2}. The gray line in Figure ~~??~~ 4 shows the resulting likelihood as a function of λ. The maximum likelihood value is $\lambda = -1.53\mathrm{Wm}^{-2}\mathrm{K}^{-1}$.

However, the simplest energy balance model ~~$\mathcal{M}_0$~~ M_0 assumes the feedback parameter is the same for climate changes in the deep past, the transient historical period, and the future. Many studies (e.g. ~~(Marvel et al., 2016; Andrews et al., 2018; Dong et al., 2020; Re~~ Marvel et al. (2016); Andrews et al. (2018); Dong et al. (2020); Rose et al. (2014); Armour et al. (2013); Gregory and Andrews (2016); M) now argue that a more appropriate model should include a "pattern effect" $\Delta\lambda$ that reflects the differences between feedbacks triggered by the observed spatial pattern of transient warming and the feedbacks expected in response to the long-term equilibrium warming pattern:

$$\Delta\lambda = \frac{\partial\lambda}{\partial T'(x)} \Delta T'(x).$$

$$M_{\Delta\lambda} : \Delta N = (\lambda - \Delta\lambda)\Delta T + \Delta F$$

~~This modifies the simple energy balance model by including a pattern effect $\Delta\lambda$:~~

$$\mathcal{M}_{\Delta\lambda} : \Delta N = (\lambda - \Delta\lambda)\Delta T + \Delta F$$

S20 placed a Gaussian prior on this pattern effect $\Delta\lambda$ =N(0.5 , 0.3) W m^{-2}K^{-1}. This corresponds to a modification of the tilt of the plane in Figure 4a. Because this model assumes the pattern effect is linearly additive, no further curvature is introduced. By multiplying the joint likelihood $P(\Delta Q, \Delta T, \Delta F|\lambda, \Delta\lambda)$ by this prior $P(\Delta\lambda)$ and integrating over all values of $\Delta\lambda$, we obtain a "marginal" likelihood for the historical evidence as a function of the feedback parameter λ. This is shown by the black line in Figure ??4b. The inclusion of the additive pattern effect and our ~~prior belief~~ physics-informed intution that it is likely to be positive shift the most likely value of the feedback parameter to $\lambda = -1.0\mathrm{Wm}^{-2}\mathrm{K}^{-1}$. ~~Likelihood for the feedback parameter λ given the simple energy balance model with no pattern effect (gray line) and marginal likelihood for λ given an additive pattern effect with prior $\Delta\lambda \sim N(0.5, 0.3)$.~~

6.2 ~~Comparing "apples to apples" when combining lines of evidence~~

~~How should we ensure that when we combine multiple lines of evidence, we can be confident that these lines are actually measuring the same thing? Bayesian methods of model evaluation provide useful checks to ensure we are indeed comparing "apples to apples".~~

~~Here is an illustrative example: suppose we firmly believe in the existence of a historical pattern effect , and we place Gaussian priors (the same as used in~~ The pattern effect estimate used in S20 ~~)~~ was based on the Atmospheric Model Intercomparison Project II (AMIPII) dataset, which produces the largest estimate of the pattern effect Modak and Mauritsen (2023), and therefore the priors on $\Delta\lambda$ used there may be both overconfident and too strongly weighted toward high values. However, while noting this important caveat, for illustrative purposes we will use the S20 historical likelihood marginalized over the pattern effect estimate as the "historical" likelihood for the rest of this paper. ~~These priors reflect our beliefs that this pattern effect acts to make future equilibrium feedbacks less negative (and thus climate sensitivity S higher).~~

~~How should we best compare these historical constraints with the evidence from the Last Glacial Maximum? The black line in all four panels of Figure 5shows the likelihood for~~

6.2 The "Twin Peaks" problem

Assuming conditional independence between lines of evidence, the posterior distribution of the feedback parameter λ ~~derived from the historical observations (with a pattern effect). In Figure 5(a), the~~ is

$$P(\lambda|Y) \propto P(Y_{hist}|\lambda)P(Y_{LGM}|\lambda)P(\lambda) \tag{7}$$

That is, the posterior estimate of λ given two lines of evidence is proportional to the product of the individual likelihoods. But what if the likelihoods have a small (or no) region of overlap? Can we really be confident that the posterior estimate is well-constrained in this case? Figure 5a highlights this potential pitfall. The black line shows the marginal likelihood for the historical evidence as a function of λ. The light blue line shows the likelihood for ~~λ derived from~~ the S20 LGM evidence as a function of λ, assuming no state dependence ~~. Why do these two distributions not overlapsubstantially? If we are confident in the evidence used, the answer must~~ ($\alpha = 0$). The product of these likelihoods is shown as a green dashed line. The less the historical and paleo likelihoods overlap, the narrower the posterior will be. We refer to this conundrum as the "Twin Peaks"

problem: should larger incompatibility between multiple lines of evidence *really* reduce the uncertainty in λ? Or could it be that the ~~LGM and historical data are~~ two lines of evidence are not, in fact, measuring the same thing?

We can take the latter possibility into account by using an alternate model for the paleo evidence. Note that the posterior for λ shown in ~~fact measuring different things, and that additional steps must be taken to ensure that the distributions reflect the same λ. The dark blue line in~~ Figure 5~~(b) shows the marginal likelihood for~~ a is conditional on a model M_0 for the paleoclimate evidence that contains only one parameter λ~~given the same LGM evidence, a model that~~ . The model assumes that the equilibrium feedbacks in a warmer climate are exactly the same as those in a colder climate, that the response to pure CO_2 forcing is equivalent to the response to LGM forcings, and that the pattern effect is zero over the LGM. An alternate model, say M_α, allows for state dependence ~~, and a Gaussian prior on the state dependence parameter~~ via an additional parameter α. The marginal likelihood for the paleo data given M_α and Gaussian priors on α is shown as a dark blue line in Figure 5b. While the overlap between these two distributions is far from exact, it is substantially larger than for the no-state-dependence case illustrated in Figure 5~~(a). Intuitively, the estimates~~ a. Simply put, the historical evidence and the LGM evidence appear to be ~~in better agreement when Last Glacial Maximum estimates of λ are assumed to be different from the historical estimates of λ.~~

more compatible when we correct for the state dependence of the past cold period. When using T20 evidence, however, there is considerable overlap between the historical (with pattern effect) and paleo (with no state dependence) likelihoods. As in the top two panels, the black lines in Figure 5c and d show the historical likelihood. The likelihood for λ obtained from T20 evidence and assuming no state dependence (orange line, Figure 5c closely overlaps the historical likelihood, as does the likelihood assuming state dependence with a prior on α as in S20 (red line, Figure 5d. The latter model, however, yields a broader likelihood for λ and therefore the region of overlap with the historical evidence is smaller.

Combining multiple lines of evidence therefore introduces another source of unavoidable subjectivity: how can we be sure that in doing so, we are comparing "apples to apples"?

6.3 Model Odds

The question of how to compare separate lines of evidence is a question of models: namely, how do we interpret the separate lines? Fortunately, Bayesian methods allow us to compare and criticize models based on the evidence. Consider, for example, two models for the LGM: M_0 and M_α. The model odds are defined as

$$\begin{aligned}\text{odds} &= \frac{\mathrm{P}(\mathrm{M}_\alpha|\mathrm{Y}_{\mathrm{hist}},\mathrm{Y}_{\mathrm{paleo}})}{\mathrm{P}(\mathrm{M}_0|\mathrm{Y}_{\mathrm{hist}},\mathrm{Y}_{\mathrm{paleo}})} \\ &= \frac{P(Y_{hist},Y_{paleo}|M_\alpha)P(M_\alpha)}{P(Y_{hist},Y_{paleo}|M_0)P(M_0)} \\ &\equiv BF \times \frac{P(M_\alpha)}{P(M_0)}\end{aligned}$$

where the Bayes Factor BF is the ratio of the evidence for each model.

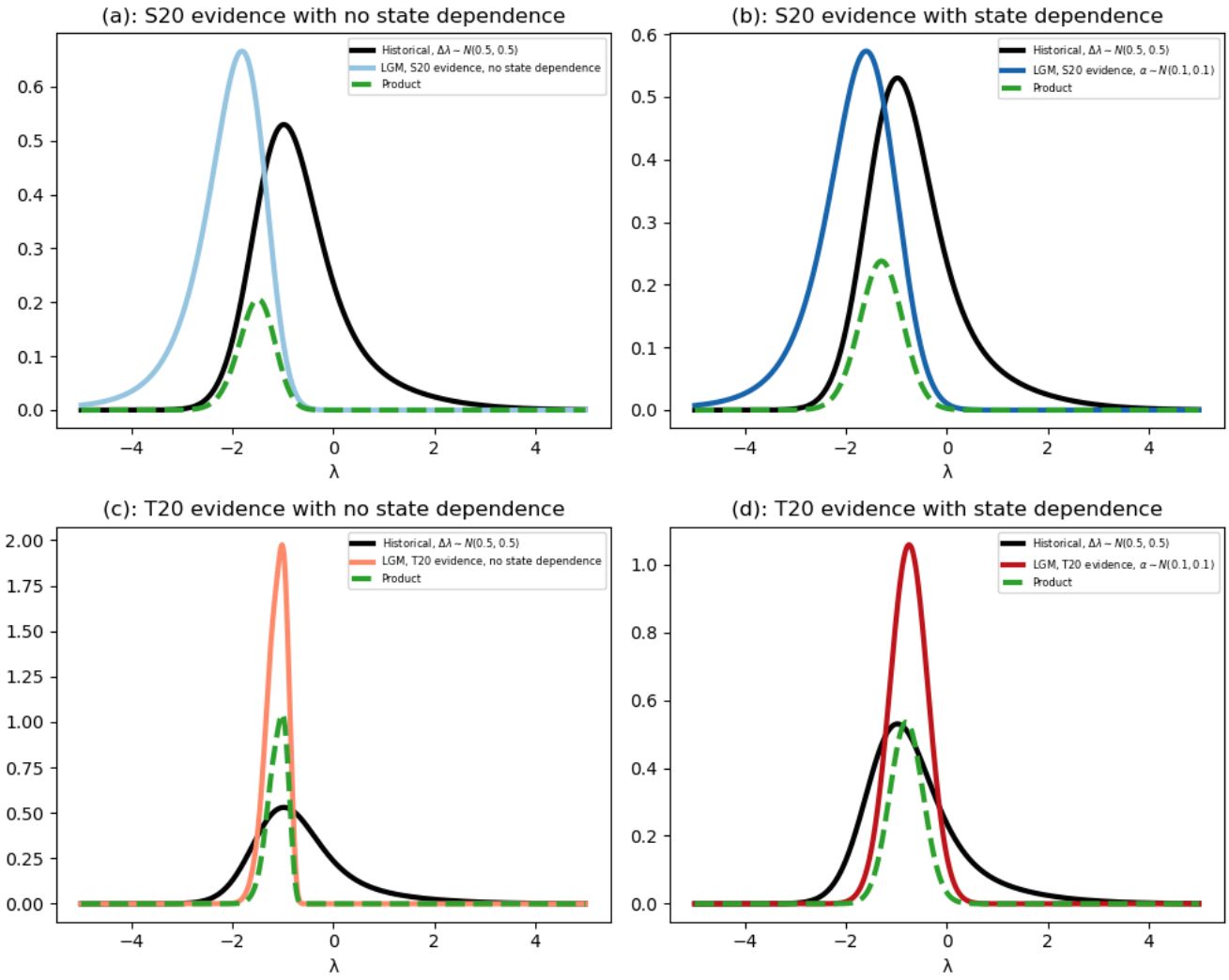

Figure 5. Likelihoods from multiple lines of evidence. In all four panels, the black line shows the likelihood for the historical evidence given λ and assuming a pattern effect $\Delta\lambda \sim N(0.5, 0.3)$. ~~(a)~~: Likelihood of S20 evidence given λ assuming no state dependence in the LGM (light blue line) and overlap (dashed green line). ~~(b)~~:Likelihood of S20 evidence given λ assuming state dependence and $\alpha \sim N(0.1, 0.1)$ (dark blue line) and overlap (dashed green line). (c): Likelihood of T20 evidence given λ assuming no state dependence in the LGM (orange line) and overlap (dashed green line). ~~(b)~~:Likelihood of T20 evidence given λ assuming state dependence and $\alpha \sim N(0.1, 0.1)$ (dark red line) and overlap (dashed green line).

The *model evidence* ~~for any given model $\mathcal{M}_\ell$ can be calculated using~~

$$P(Y|\mathcal{M}_\ell) = \int P(Y|\Theta, \mathcal{M}_\ell) P(\Theta|\mathcal{M}_\ell) d\Theta.$$

for any given model M_ℓ is defined as the integrated likelihood over all values of its parameters Θ_ℓ:

$$P(Y|M_\ell) = \int P(Y|\Theta, M_\ell) P(\Theta_\ell|M_\ell) d\Theta_\ell. \quad (8)$$

~~This reflects the probability that model $\mathcal{M}_\ell$~~ This reflects the probability that model M_ℓ could have generated the observed evidence ~~under a given set of prior beliefs about its parameters θ. The ratio of evidence for two models $\mathcal{M}_i$ and $\mathcal{M}_j$ defines the Bayes factor (BF) , which updates the prior odds in favor or against a model. We note that the model evidence depends both on the model used and on the priors for each parameter in the model. This has the useful property of penalizing models with many parameters that do not add value in fitting the evidence, thereby avoiding over-fitting.~~ under a given set of priors on its parameters θ_ℓ.

~~Consider, for example, comparing the model used in Figure 5a to the model used in Figure 5b. In panel (a), the modelassumes no state dependence for the LGM evidence but a pattern effect in the transient historical evidence ; we will denote this as $\mathcal{M}_{0,\Delta\lambda}$. In panel (b), the model assumes the feedbacks depend on the background temperature and places the Gaussian prior used in S20 on the parameter α, which we denote as $\mathcal{M}_{\alpha,\Delta\lambda}$. Using Eq. 8, the evidence for the first model $\mathcal{M}_{0,\Delta\lambda}$ is given by~~

$$P(Y_{hist}, Y_{paleo}|\mathcal{M}_{0,\Delta\lambda}) \propto \int\int P(Y_{paleo}|\lambda) P(Y_{hist}|\lambda, \Delta\lambda) P(\lambda) P(\Delta\lambda) d\lambda\, d(\Delta\lambda)$$

For example, the model evidence for model M_0 is

$$P(Y_{hist}, Y_{paleo}|M_0) \propto \int P(Y_{paleo}|\lambda) P_{\Delta\lambda}(Y_{hist}|\lambda) P(\lambda) d\lambda$$

~~The terms in the integrand are, from left to right: the likelihood of the paleo evidence given λ, the likelihood of~~ where $P_{\Delta\lambda}(Y_{hist}|\lambda)$ is the marginal historical likelihood (black line, Figure 5a). When combined with a uniform prior on λ, the ~~historical evidence given λ and $\Delta\lambda$, and priors on the feedbacks λ (here, assumed uniform on -10, 10) and the pattern effect $\Delta\lambda$ (as in S20, assumed to be N(0.5,0.3))~~model evidence for M_0 is therefore the area under the green curve in Figure 5a.

~~The~~ By contrast, the model evidence for the ~~second model is~~ model M_α is

$$P(Y_{hist}, Y_{paleo}|\mathcal{M}_{\alpha,\Delta\lambda} M_\alpha) \propto \int\int P(Y_{paleo}|\lambda, \alpha) P_{\Delta\lambda}(Y_{hist}|\lambda, \Delta\lambda) P(\lambda\alpha) P(\Delta\lambda) P(\alpha) d\lambda\alpha\, d(\Delta\lambda)\, d\alpha$$

~~We can then calculate the Bayes factor~~

$$BF = \frac{P(Y|\mathcal{M}_{\alpha,\Delta\lambda})}{P(Y|\mathcal{M}_{0,\Delta\lambda})} = \frac{P(Y_{hist}, Y_{paleo}|\mathcal{M}_{\alpha,\Delta\lambda})}{P(Y_{hist}, Y_{paleo}|\mathcal{M}_{0,\Delta\lambda})} = 1.33.$$

~~If our prior belief~~ When combined with a uniform prior on λ, the model evidence for M_α is the area under the green curve in Figure 5b.

Using S20 evidence and these priors, we find that the Bayes Factor is 1.33. This means that if our prior is that both models are equally likely, the evidence shifts those odds: the model depicted in panel ~~(b)~~ b is about 33% more likely to have generated the observed paleo and historical evidence.

~~Does this mean that the evidenceindicates the definite existence of state dependence in the paleoclimate data? Certainly not. It simply means that given the LGM and historical evidence used in S20 assuming the only candidate models to interpret the LGM evidenceare those with and~~ However, using T20 evidence, the Bayes factor is 0.93. This suggests that the "better" model to use, given T20 evidence, is one without state dependence~~and using S20's Gaussian priors on α that this is are the model that maximizes the agreement between separate lines of evidence. We are *not* arguing that this is the objectively "correct~~. Clearly, the "best" ~~way to combine the Last Glacial Maximum reconstructions with historical observations.~~

~~In fact, whether or not~~ model depends on the evidence used, the prior knowledge of whether we are comparing "apples to apples"~~depends very heavily on the evidence we use. As in the top two panels, the black lines in Figure 5(c) and (d) show the historical likelihood assuming a pattern effect with S20's Gaussian prior. The likelihood for λ obtained from T20 evidence and assuming no state dependence (orange line, Figure 5(c)) closely overlaps the historical likelihood, as does the likelihood assuming state dependence with a prior on α as in S20 (red line, Figure 5(c)). The latter model, however, yields a broader likelihood for λ and therefore the region of overlap with the historical evidence is smaller. The Bayes factor using T20 evidence is calculated as~~

$$BF = \frac{P(Y|\mathcal{M}_{\alpha,\Delta\lambda})}{P(Y|\mathcal{M}_{0,\Delta\lambda}} = 0.93.$$

~~This suggests that~~ , and the ~~"better" model to use, given T20 evidence, is one without state dependence.~~

7 ~~A Way Forward~~

~~In the sections above, we have demonstrated that constraints on the feedback parameter~~ priors we place on λ~~depend heavily on the evidence~~ , $\Delta\lambda$, and α.

We note that whether the twin peaks problem is indeed a "problem" is largely dependent on the prior odds $P(M_\alpha)/P(M_0)$, which must be specified. If we have prior knowlege that the ~~model used to interpret that evidence, and prior beliefs about the parameters. Moreover, our ability to compare different~~ two lines of evidence are measuring the same thing, then we will give more prior weight to the simple model M_0 and the Bayes Factor will do little to shift the odds. This will result in a narrower posterior estimate: if two lines of evidence ~~also depend on the evidence, models and priors we use to do so. It might initially appear that we are doomed to wallow in subjectivity, with no hope of arriving at credible, usable estimates of λ or S. However, it is possible to move forward by relying on a community of experts, all of whom must be willing to clearly specify their prior beliefs and update their understandings in light of evidence~~are compatible only for a small range of values, and we are confident in what the evidence is telling us, then we may be more confident in its posterior value.

~~Any expert assessment of S, λ, or indeed any other climate-relevant parameter (e.g. the Zero Emissions Commitment or the Transient Climate Response) should: Easily update estimates as new information comes in Compare "apples to apples" when combining~~

7 A Way Forward

Thus far, we have established that there are three places where unavoidable subjective decisions must be made: collecting evidence, choosing the interpretive model, and assessing prior knowledge of that model's parameters. We have also established that multiple lines of evidence ~~. Handle differing expert opinion in a fair and systematic way.~~ appear more or less compatible depending on the models used. Here, we present a suggested ~~way forward for expert assessment . Every analysis will require subjective decisions; we seek to both make these decision points explicit and allow for the fair aggregation of different expert choices.~~ framework for making these decisions in a community assessment framework.

7.1 ~~Assessing evidence uncertainty~~

7.1 Handling evidence uncertainty

~~Bayesian methods are useful because they easily allow for hierarchical modeling, in which we can formulate sub-models to account for information on multiple levels, and easily propagate uncertainties. One of the most useful applications of hierarchical modeling is Bayesian~~ Whether and how much a newly published estimate of a particular quantity (for example, ΔT or ΔF from the Last Glacial Maximum) affects the evidence base depends on prior knowledge of that quantity. It also depends on expert assessment of how the new study relates to existing literature. A single highly certain, high-quality study can strongly shift previously uncertain estimates, while low-quality or uncertain published estimates may not change previously firm understandings.

We suggest formalizing these intuitions using a Bayesian random effects meta-analysis ~~(Smith et al., 1995)~~Smith et al. (1995) , frequently used in fields as diverse as psychology ~~(Gronau et al., 2021), medicine (Sutton and Abrams, 2001)~~Gronau et al. (2021) , medicine Sutton and Abrams (2001), and ecology~~(Koricheva et al., 2013). To combine multiple studies, we assume:~~ Koricheva et al. (2013 . This model can be written as

$$y_i \hat{y}_j \sim \mathcal{N}\mathcal{N}(\theta_i y_j, \sigma_{ij}) \tag{9}$$

$$\theta_i y_j \sim \mathcal{N}\mathcal{N}(\mu Y, \tau) \mu \sim g(.) \tau \sim h(.) \tag{10}$$

~~Here y_i and σ_i~~ where $\hat{y}_j$ and σ_j are the reported mean and ~~uncertainty of the evidence (i.e., ΔT, ΔN, or ΔF) from study i. The reported mean y_i is assumed to be distributed about a some~~ standard deviation of each study j. We assume the true (latent) ~~value θ_i. This parameter depends on the study i, and can roughly be thought of as the value around which all subsequent repetitions of the work would be distributed. Each of these study means θ_i are then assumed to be drawn from a distribution with common mean μ and inter-study spread~~ mean y_j of each study is normally distributed about an overall mean Y, with τ ~~.~~

~~The priors $g(.)$ on μ and $h(.)$ on τ reflect our prior beliefs about both the true value of the parameter μ and the design of the studies. In a "fixed effects" meta-analysis the~~ the expected inter-study standard deviation.

The priors we put on the quantities of interest– the overall mean Y and the between-study spread τ ~~is assumed to be zero . This means that all reported estimates y_i share a common mean, andany differences are simply due to sampling error~~– quantify our previous knowledge of and views about the literature. A τ very close to zero suggests homogeneity across studies (and, in fact, choosing to set $\tau = 0$ reduces the random-effects model to the fixed-effects model). By contrast, ~~in a "random effects" meta-analysis there are assumed to be structural differences between individual studies that mean that we should expect variation between estimates~~if we have reason to believe that multiple studies should vary in their reported values due to structural and design factors, then we might place a broad prior on τ. For example, a fixed-effects model might be appropriate for calculating the ensemble mean of a quantity within a single CMIP model, whereas a random-effects model might be more appropriate for combining ensembles of multiple CMIP models, which we know to differ structurally.

As ~~an~~ a specific example relevant to calculating the feedback parameter λ, ~~we will~~ consider multiple published LGM global mean temperature changes ΔT derived from proxies and models as well as from PMIP3 and PMIP4 models (Table 1).

~~How best should we combine these estimates to obtain a single distribution of ΔT and its uncertainty? This depends on many things: the literature on which we rely, our judgement about how best to pool multiple estimates into a single uncertain quantity, and the reported quantities in the studies themselves. How cold was the Last Glacial Maximum? The answer depends on your prior beliefs about the cooling and about the literature. Shown are posterior distributions for the LGM cooling ΔT assuming a random effects model and broad (blue line) or T20 (green) priors on the mean or a fixed effects model and broad (orange line) or T20 (red line) priors on the mean.~~

Figure 6 illustrates how the posterior distribution of ΔT depends on prior beliefs about the nature and quality of the published literature assessing it. Consider, for example, a random-effects model in which we place broad priors on the mean $\mu \sim N(0, 100)$ and inter-study standard deviation $\tau \sim U(0, 100)$. With these prior assumptions, 90% of the resulting posterior density for μ (the true value of ΔT) lies between (-5.9K, -4.8K). Assuming that there is *no* inter-study spread (i.e, τ is assumed to be zero with zero uncertainty: a fixed effect model) would yield an estimate of ΔT 90% likely to be between -4.8 and -4.5K. This much narrower (and warmer) estimate results from the extremely restrictive prior belief that every study, regardless of method, targets the same underlying ΔT and would yield the same results if performed perfectly and with adequate data. Similarly, ~~if we believe the~~ we might set the prior on μ using the result of a single published study (say, for example~~T20, to be accurate, then we may adopt this as our prior belief, setting the prior on μ to be the T20 distribution of~~ , ΔT from T20). Combined with a broad uniform prior on the inter-study spread, this results in an 90% posterior density estimate of (-6.2K, -5.6K). If, however, we adopt the restrictive fixed effects model, the T20 study is merely treated as an outlier and fails to substantially move the posterior distribution toward cooler values of ΔT (red line), even ~~if our prior belief is that T20 is exactly correct.~~

~~What this means is that a simple hierarchical model for the evidence (ΔT or ΔF, for example) allows different experts to specify their priorbeliefs about the true values of the evidence and the literature. Their resulting posteriors will depend strongly on those prior assessments. This simply reflects the fact that different experts give different weights to studies in the published~~

Mean (K)	Standard Deviation	Reference	Derived From
-4.00	0.41	~~(Annan and Hargreaves, 2013)~~ Annan and Hargreaves (2013)	Proxies and models
-5.80	0.77	~~(von Deimling et al., 2006)~~ von Deimling et al. (2006)	Proxies and models
-6.20	0.46	~~(Holden et al., 2009)~~ Holden et al. (2009)	GENIE-1
-3.58	0.12	~~(Shakun et al., 2012)~~ Shakun et al. (2012)	Proxies
-6.20	0.92	~~(Snyder, 2016)~~ Snyder (2016)	Proxies and models
-6.30	0.61	~~(Bereiter et al., 2018)~~ Bereiter et al. (2018)	Proxies (ocean temperature) and models
-5.70	0.20	~~(Friedrich and Timmermann, 2020)~~ Friedrich and Timmermann (2020)	N/A
-5.75	0.38	~~(Friedrich et al., 2016)~~ Friedrich et al. (2016)	SST proxies and a model simulation
-6.10	0.20	~~(Tierney et al., 2020)~~ Tierney et al. (2020)	proxies and isotope-enabled climate mod
-5.00	1.00	~~(Sherwood et al., 2020)~~ Sherwood et al. (2020)	Synthesis
-4.85	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	CESM
-2.70	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	CNRM
-4.63	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	FGOALS-g2
-4.92	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	GISSE2-p1
-5.19	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	GISSE2-p2
-4.64	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	IPSL
-5.40	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MIROC
-4.41	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MPI-p1
-4.67	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MPI-p2
-4.71	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MRI
-3.75	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	AWIESM1
-3.81	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	AWIESM2
-6.80	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	CESM1-2
-7.16	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	HadCM3-PMIP3
-5.92	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	HadCM3-ICE6GC
-6.46	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	HadCM3-GLAC1D
-3.28	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	iLOVECLIM-ICE-6G
-3.26	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	iLOVECLIM-GLAC1D
-3.73	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	INM-CM4-8
-4.63	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	IPSLCM5A2
-4.02	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MIROC-ES2L
-3.90	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	MPI-PMIP4
-5.27	N/A	~~(Kageyama et al., 2021)~~ Kageyama et al. (2021)	UT-CCSM4

Table 1. Estimates of global cooling ΔT during the Last Glacial Maximum

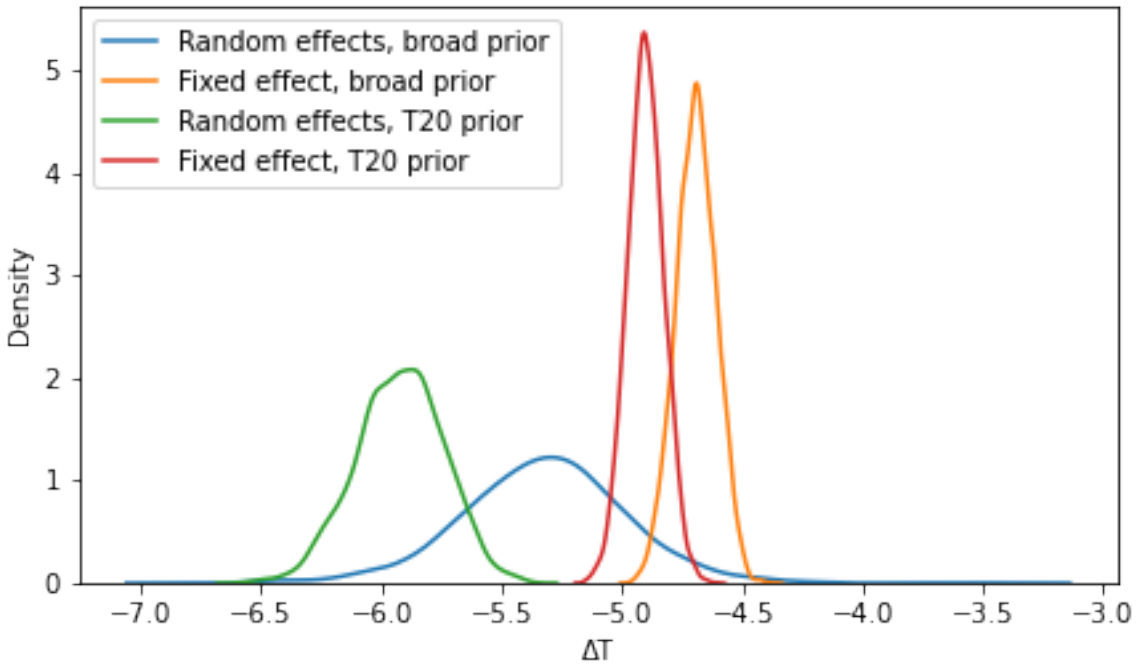

Figure 6. How cold was the Last Glacial Maximum? The answer depends on your prior beliefs about the cooling and about the literature. Shown are posterior distributions for the LGM cooling ΔT assuming a random effects model and broad (blue line) or T20 (green) priors on the mean or a fixed effects model and broad (orange line) or T20 (red line) priors on the mean.

~~literature. It is a mathematical expression of the subjective expert judgment inherent in science. We propose a method for aggregating expert judgement in section 7.3.~~

~~To obtain new constraints on λ from a meta-analysis of published LGM temperature estimates, we adopt broad priors on μ and τ (blue line, Figure 6). We also perform a similar meta-analysis for the radiative forcing using the values in Table ?? (assuming a fixed-effects model ($\tau = 0$) in the absence of reported standard deviations and broad priors N(0,100) on all means), which results in an estimate of $\Delta F = N(-8.1, 1.5)$ Wm^{-2}. Using these constraints and assuming no state dependence, the most likely value is $\lambda = -1.54$ with 5-95% range of (-2.09, -1.08)$\mathrm{Wm}^{-2}\mathrm{K}^{-1}$, corresponding to a 5-95% range of (1.87, 3.75)K for S. We will explore the impact of different models and~~ using the T20 prior.

7.1.1 Recommendations

Unavoidable subjective decisions about the *evidence* can be made explicit by adopting a random effects meta-analysis. This requires the specification of priors ~~on this estimate in sections 7.2 and 7.3, respectively~~on the inter-study spread τ and the

overall mean Y. Our recommendation is that the organizers of community assessments choose and clearly specify these priors, rather than allowing individual experts to choose their own.

~~Quantity Mean (Wm^{-2}) Standard Deviation Reference Model Generation ΔF_{ice} -3.79 N/A (Braconnot and Kageyama, 2015) CCSM4 PMIP3 ΔF_{ice} -4.90 N/A (Braconnot and Kageyama, 2015) IPSL-CM5A-LR PMIP3 ΔF_{ice} -5.20 N/A (Braconnot and Kageyama, MIROC-ESM PMIP3 ΔF_{ice} -4.57 N/A (Braconnot and Kageyama, 2015) MPI-ESM-P PMIP3 ΔF_{ice} -3.62 N/A (Braconnot and Kageyam MRI-CGCM3 PMIP3 ΔF_{ice} -2.59 N/A (Braconnot et al., 2012) CCSM3 PMIP4 ΔF_{ice} -2.66 N/A (Braconnot et al., 2012) CMRM PMIP4 ΔF_{ice} -3.23 N/A (Braconnot et al., 2012) HadCM3M2 PMIP4 ΔF_{ice} -3.41 N/A (Braconnot et al., 2012) HadCM3M2 v PMIP4 ΔF_{ice} -3.48 N/A (Braconnot et al., 2012) IPSL-CM4 PMIP4 ΔF_{ice} -2.88 N/A (Braconnot et al., 2012) MIROC3.2 PMIP4 ΔF_{ice} -3.29 N/A (Tierney et al., 2020) CESM1.2 PMIP4 ΔF_{GHG} -2.60 0.10 (Zhu and Poulsen, 2021) N/A N/A ΔF_{GHG} -2.48 0.15 (Tierney et al., 2020) N/A N/A ΔF_{GHG} -3.15 0.26 (Sherwood et al., 2020) N/A N/A ΔF_{Dust} -0.12 N/A (Albani et al., 2014); (Albani and Mahowald, 2019) C4fn-lgm N/A ΔF_{Dust} -0.36 N/A (Hopcroft et al., 2015)HadGEM2-A fixPIveg N/A ΔF_{Dust} -1.10 N/A (Hopcroft et al., 2015) HadGEM2-A N/A ΔF_{Dust} -0.32 N/A (Hopcroft et al., 2015) HadGEM2-A-DEAD N/A ΔF_{Dust} -2.00 N/A (Claquin et al., 2003) Exp1 (ext. mixing) N/A ΔF_{Dust} -1.00 N/A (Claquin et al., 2003) Exp2 (int. mix. Hem.) N/A ΔF_{Dust} -0.48 N/A (Mahowald et al., 2006) SOMB / SOMBLGMT N/A ΔF_{Dust} -0.01 N/A (Takemura et al., 2009) N/A N/A ΔF_{Dust} 0.10 N/A (Yue et al., 2011) PRND / LGM.DST N/A $\Delta F_{insolation}$ 0.01 N/A (Braconnot and Kageyama, 2015) CCSM4 N/A $\Delta F_{insolation}$ 0.01 N/A (Braconnot and Kageyama, 2015)IPSL-CM5A-LR N/A $\Delta F_{insolation}$ 0.13 N/A (Braconnot and Kagey MIROC-ESM N/A $\Delta F_{insolation}$ 0.01 N/A (Braconnot and Kageyama, 2015)MPI-ESM-P N/A $\Delta F_{insolation}$ 0.01 N/A (Braconnot and Kag MRI-CGCM3 N/A $\Delta F_{vegetation}$ -1.1 0.6 (Köhler et al., 2010) N/A N/A Estimates of ΔF for the Last Glacial Maximum from ice sheets, solar insolation, dust, and vegetation.~~

7.2 Handling model uncertainty

As shown in Section 4, the constraints placed on climate sensitivity by ~~paleo or historical evidence~~ multiple lines of evidence evidence depend on the model(s) used to interpret that evidence. This means that the design of every expert assessment must be explicit about ~~the modelsused to interpret each line of evidence~~its interpretive models. As the assessment is planned, it is crucial to arrive at consensus on credible interpretive models for the evidence. For example, one possible model for the Last Glacial Maximum might incorporate parameters α (representing state dependence), ξ (representing the difference between long-term equilibrium LGM feedbacks and the target quasi-equilibrium feedbacks to doubled CO2) and $\Delta\lambda_{LGM}$ (representing radiatively important sea-surface pattern differences between the LGM and doubled CO2):

$$\Delta T = \frac{-\Delta F}{\frac{\lambda+\Delta\lambda_{LGM}}{1+\xi} + \frac{\alpha}{2}\Delta T}$$

Given a model, ~~even an unwieldy one with multiple parameters,~~ experts may then be asked to specify their prior beliefs about each parameter. If an expert disagrees with the inclusion of a parameter in a model, s/he would be free to set a prior very narrowly clustered around 0 on that prior.

If consensus cannot be reached on a particular model, then we suggest that the planning team for any assessment arrive at a list of candidate models ~~$\mathcal{M}_1 \ldots \mathcal{M}_K$~~ $M_1 \ldots M_K$. The aggregate posterior can then be taken as a weighted average over different

models:

$$P(\Theta|Y) = \sum_{k=1}^{K} w_k P(\Theta|\mathcal{M}M_k, Y). \tag{11}$$

Here, $(\Theta|M_k, Y)$ is the posterior obtained using the model ~~$\mathcal{M}_k$~~ M_k to interpret the evidence Y.

The weights reflect how well the model fits the data, and are given by

$$w_k = P(\mathcal{M}M_k|Y) = \frac{P(Y|\mathcal{M}_k)P(\mathcal{M}_k)}{\sum_{k=1}^{K} P(Y|\mathcal{M}_k)P(\mathcal{M}_k)} \frac{P(Y|M_k)P(M_k)}{\sum_{k=1}^{K} P(Y|M_k)P(M_k)}. \tag{12}$$

The term ~~$P(\mathcal{M}_k|Y)$~~ $P(M_k|Y)$ is the model evidence (Eq 8, discussed in section 6.2). These weights, and hence the combined posterior, depend on the priors ~~$P(\mathcal{M}_k)$~~ $P(M_k)$ we put on the correctness of each model. If an assessment allows for experts to use one of multiple models, it is therefore imperative to specify assessment-wide priors on these models upfront.

~~As a worked example, consider two models: $\mathcal{M}_{0,\Delta\lambda}$ in which the paleoclimate evidenceis assumed to have no state dependence but a pattern effect is present in the historical observations, and $\mathcal{M}_{\alpha,0}$ in which there is no pattern effect in the historical observations but we allow for state dependence in the LGM. Table ?? shows the resulting estimates given different prior beliefs about these models. In all cases, the prior on λ is assumed to be U(-10,10) and the prior on all parameters is as given in S20.~~

7.2.1 Recommendations

We recommend that organizers of community assessments clearly specify a single interpretive model for the evidence. If this is not possible, organizers should specify a list of possible candidate models M_k and ask and a prior $P(M_k)$ for each candidate model. The resulting estimate will then be a weighted averages over the models.

~~Model Priors $P(\Theta)$ $\lambda(95\%CL)$ $S(95\%CL)$ $P(\mathcal{M}_{0,\Delta\lambda}) = 1$ $P(\Delta\lambda) = N(0.5, 0.3)$ (-1.9,-1.1) (2.1,4.0) $P(\mathcal{M}_{\alpha,0}) = 1$ $P(\alpha) = N(0.1, 0.$ (-1.9,-0.7) (2.1,5.4) $P(\mathcal{M}_{0,\Delta\lambda}) = P(\mathcal{M}_{\alpha,0}) = 0.5$ as above (-1,9,-0.8) (2.1,5.0) Estimates of feedbacks and climate sensitivity given different prior assumptions about models.~~

7.3 Expert elicitation via priors

Finally, it is necessary to quantify the degree of pre-existing knowledge and/or beliefs through the use of prior distributions. ~~These enter the present analysis in three different places: first, in beliefs about the multiple studies used to constrain the evidence (priors on μ and τ in Section 7.1), second in beliefs about the underlying model used to explain the evidence and finally, in beliefs about the distributions of the parameters θ in those models.~~ This is where a wide variety of expert opinion may be usefully incorporated in an assessment.

However, we require consistent ways to aggregate the judgement of multiple experts. In theory, sufficient evidence should lead to a high degree of agreement~~among experts~~, even if ~~they~~ different experts begin the analysis with ~~different prior beliefs~~very different priors. Figure 7~~(a) shows the prior beliefs of~~ a shows the priors placed on the parameter λ by two hypothetical experts. Expert A (solid red line) believes the feedback parameter to be less negative than Expert B (solid blue line) and is even open to the idea that it might be positive. Dashed red and blue lines show both experts' posteriors, when updated using the evidence

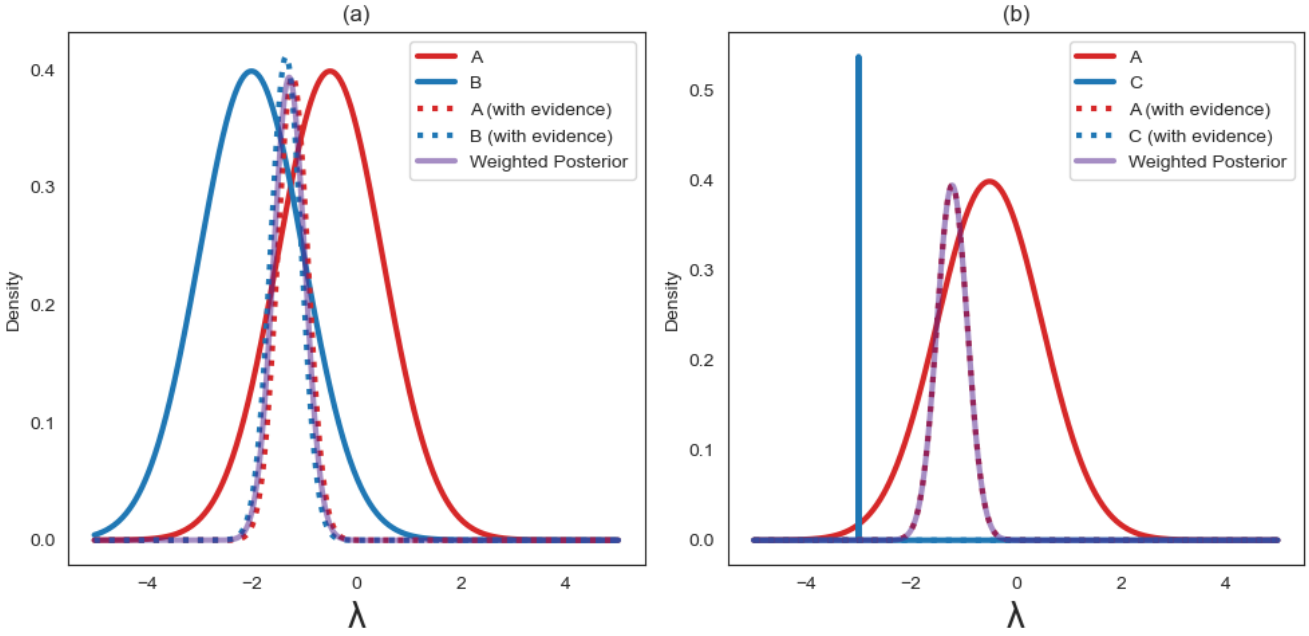

Figure 7. ~~(a)~~: Experts A (solid red line) and B (solid blue line) begin with different priors on λ. The evidence presented in S20 updates these priors, and the resulting posteriors are nearly identical (dotted red and blue lines). The purple line shows the weighted posterior. ~~(a)~~: Experts A (solid red line) and C (solid blue line) begin with different priors on λ, but C's prior is very narrowly peaked. The evidence presented in S20 updates these priors, but the posteriors remain very different(dotted red and blue lines). The purple line shows the weighted posterior, which is almost identical to A's posterior.

presented in S20. While the experts began their analysis with differing opinions, the weight of the evidence has updated their understandings and they now agree about the feedback parameter λ. However, some experts may not be as open-minded as our researchers A and B. Expert C (blue line, Figure 7~~(b)~~ b believes the feedback parameter to be strongly negative. Moreover, s/he is extremely confident in this: his~~prior distributions~~ /her prior distribution is very narrowly peaked around a value of $\lambda = -3Wm^{-2}K^{-1}$. Expert C's confidence remains unshaken by the evidence presented in S20, and his/her posterior remains nearly identical to his/her prior beliefs. How should an assessment handle such excessively confident experts, whose beliefs appear to be unshakeable by any reasonable amount of evidence?

Consider an assessment in which N experts each specify their priors $P_i(\theta)$, where $i = 1 \ldots N$. A reasonable aggregate prior might then be a linear combination of the individual expert priors:

$$P(\theta) = \sum_{i=1}^{N} a_i P_i(\theta).$$

$$P(\theta) = \sum_{i=1}^{N} a_i P_i(\theta). \tag{13}$$

The aggregate posterior is therefore a weighted average of the individual expert posteriors

$$P(\theta, Y) = \sum_i \tilde{a}_i P_i(\theta|i, Y)$$

$$P(\theta, Y) = \sum_i \tilde{a}_i P_i(\theta|i, Y) \tag{14}$$

where

$$\tilde{a}_i = \frac{a_i \int P(Y|\theta) P_i(\theta) d\theta}{\sum_{i=1}^{N} a_i \int P(Y|\theta) P_i(\theta) d\theta}.$$

$$\tilde{a}_i = \frac{a_i \int P(Y|\theta) P_i(\theta) d\theta}{\sum_{i=1}^{N} a_i \int P(Y|\theta) P_i(\theta) d\theta}. \tag{15}$$

This method introduces N new parameters: the prior weight a_i we assign to each expert's judgement. This is a far easier task than setting priors on models (as discussed in Section 7.2) because it requires no physical understanding, only a belief about the "quality" of each expert's initial beliefs. We recommend weighting each expert equally by setting $a_1 = a_2 = \ldots a_N = \frac{1}{N}$, in which case the posterior weights become

$$\tilde{a}_i = \frac{\int P(Y|\theta) P_i(\theta) d\theta}{\sum_{i=1}^{N} \int P(Y|\theta) P_i(\theta) d\theta}.$$

$$\tilde{a}_i = \frac{\int P(Y|\theta) P_i(\theta) d\theta}{\sum_{i=1}^{N} \int P(Y|\theta) P_i(\theta) d\theta}. \tag{16}$$

The purple line in Figure 7~~(a)~~ a shows the resulting aggregate posterior given A and B's priors. Because both these experts are similarly able to update their priors, the weighting process has no effect on the outcome. However, the weighted average of A and C's posteriors, shown as a purple line in 7~~(b)~~b, is similar to A's posterior distribution. The narrowness of C's prior causes his/her posterior distribution to be down-weighted in the weighted average. We suggest this as an effective strategy for handling inflexible or extremely anomalous expert opinions.

8 ~~Discussion and Conclusions~~

7.0.1 Recommendations

We recommend eliciting expert judgement in a systematic way by allowing experts to specify priors on pre-determined model parameters. The analysis can then be performed using a single aggregate posterior calculated as the weighted average of individual expert posteriors.

8 Conclusions

Here, we have presented three sources of uncertainty that enter in to estimates of climate sensitivity. First, what evidence are we using to constrain climate sensitivity, ~~where do those estimates or measurements come from~~how do we decide what counts as "evidence", and how should we handle estimates that disagree or conflict? Second, what interpretive model should we be using to relate the evidence to the climate sensitivity, and what parameters are required? Third, what prior knowledge of these parameters is it appropriate to include? ~~In the subsequent section, we have laid out the rudiments of~~ We then propose a strategy to ~~combine multiple published estimates of variables relevant to climate sensitivity~~make the role of expert judgement in subsequent assessments fairer and more transparent. The advantage of this strategy, combining Bayesian meta-analysis and Bayesian model averaging, is that it can incorporate newly published data and is easily expanded to handle uncertainties at multiple levels.

There is no limit to the number of nested levels we could theoretically use within a Bayesian hierarchical model: the prior for radiative forcing from ice sheets, for example, can be updated using a global ice sheet reconstruction, which itself is constrained by individual geological measurements. Similarly, a prior on ocean heat uptake ΔN or historical warming ΔT can be updated as new measurements become available. However, to remain tractable every project must truncate the hierarchy at some finite level. In practice, this means treating the posteriors that arise from observational, GCM, or paleoclimate studies as evidence; where we draw the line between evidence and parameter sets the bounds of our analysis.

As a result, we propose a framework in which experts are required to specify their choices at clearly defined decision points. Once priors are specified, the model and evidence will update them accordingly, arriving at a new, aggregate consensus posterior. We review this framework here.

Somewhat obviously, experts' beliefs about the data are based on their prior beliefs, updated by the evidence. But how they interpret and use that evidence depends on the subjective choices they make: what counts as a "study" or "evidence"? How should we best compare estimates derived from proxies or observations and estimates from GCMs? Should some studies receive more weight than others? In our framework, experts must make ~~make~~ the following judements about the evidence:

1. What is your informed belief about the evidence? (E.g. what is your prior on μ?)

2. What is your belief about the published literature? (What is your prior on τ)

Second, we suggest taking the choice of model out of individual participants' hands to the greatest extent possible. Ideally, assessment planners would arrive at a single model and set of parameters on which experts may specify their priors. If not, they should arrive at a list of candidate models, specify firm prior beliefs about these models, and perform Bayesian model averaging over the posteriors of individual experts, which will depend on the model they use.

Third, once a model is specified, experts should specify their prior beliefs about the parameters of that model.

The results presented here are meant to begin, not end, a conversation. The beauty of Bayesian methods is that we can allow new evidence to update our existing ~~beliefs~~knowledge. As climate researchers gear up for the next generation of model intercomparison projects and assessments, it is important to consider how these new results will be integrated with existing knowledge. Our methods presented here allow for new discoveries to advance our understanding, ultimately narrowing the bounds of climate sensitivity and informing future research and decision making.

Code availability. TEXT

Data availability. TEXT

Code and data availability. TEXT

Sample availability. TEXT

Video supplement. TEXT

Appendix A: ~~1~~Exact forms of integrals

A1 ~~Exact forms of integrals~~

To estimate the likelihood of the evidence ΔT and ΔF given the simple energy balance model, we integrate the joint probability distribution $\mathcal{J}(\Delta T, \Delta F)$ over the curve C defined by the model :

$$P(Y|\lambda, \mathcal{M}M_0) = \int_C \mathcal{J}(\Delta T, \Delta F) ds \tag{A1}$$

~~The curve~~ C can be parameterized as

$$\mathbf{r}(t) = t\hat{i} + -\lambda t\hat{j}$$

$$\mathbf{r}(t) = t\hat{i} + -\lambda t\hat{j} \tag{A2}$$

and the integral is then

$$P(Y|\lambda, \mathcal{M}M_0) = \int_{-\infty}^{\infty} \mathcal{J}(\mathbf{r}(t))\,||\mathbf{r}'(t)||\,dt = \int_{-\infty}^{\infty} \mathcal{J}(t, -\lambda t)\sqrt{1+\lambda^2}\,dt. \tag{A3}$$

In the case where ΔT and ΔF are Gaussian and independent with means μ_T, μ_F and standard deviations σ_T, σ_F respectively, the likelihood has an exact analytic form, substantially speeding up its computation:

$$P(Y|\lambda, \mathcal{M}M_0) = C\left(\frac{2\pi}{A}\right)^{1/2} \exp\left(\frac{B^2}{2A}\right) \tag{A4}$$

where

$$\begin{aligned} C &= \frac{\sqrt{1+\lambda^2}}{2\pi\sigma_T\sigma_F} \exp\left(\frac{\mu_T^2}{\sigma_T^2} + \frac{\mu_F^2}{\sigma_F^2}\right) \\ A &= \frac{1}{\sigma_T^2} + \frac{\lambda^2}{\sigma_F^2} \\ B &= \frac{\mu_T}{\sigma_T^2} - \frac{\lambda\mu_F}{\sigma_F^2} \end{aligned}$$

In the case of a three-dimensional space (as for the historical evidence), the curve C defines a plane, not a line, and we have

$$P(Y|\lambda) \propto \int_C \mathcal{J}(\Delta T, \Delta F, \Delta N)dS = \int\int \mathcal{J}\left(\boldsymbol{r}(u,v)\right)||r_u \times r_v||du\,dv$$

$$P(Y|\lambda) \propto \int_C \mathcal{J}(\Delta T, \Delta F, \Delta N)dS = \int\int \mathcal{J}\left(\boldsymbol{r}(u,v)\right)||r_u \times r_v||du\,dv \tag{A5}$$

where

$$\boldsymbol{r} = u\hat{i} + v\hat{j} + (\lambda u + v)\hat{k}$$

$$\boldsymbol{r} = u\hat{i} + v\hat{j} + (\lambda u + v)\hat{k} \tag{A6}$$

~~In the case where the uncertainties in temperature, energy imbalance, and historical forcing are all Gaussian and uncorrelated, the likelihood can be evaluated analytically:~~

$$\mathcal{L}(\lambda_{hist}|\Delta T, \Delta F, \Delta N) = 2\pi\sqrt{\frac{\lambda_{hist}^2+2}{\det(A)}}\exp\left\{\frac{1}{2}\left(J^T A^{-1} J - C\right)\right\}$$

~~where~~

$$A = \begin{bmatrix} \sigma_T^{-2} + \lambda^2\sigma_N^{-2} & \lambda\sigma_N^{-2} \\ \lambda\sigma_N^{-2} & \sigma_F^{-2} + \sigma_N^{-2} \end{bmatrix}$$

$$J = \begin{bmatrix} \mu_T^2\sigma_T^{-2} + \lambda_{hist}\mu_N\sigma_N^{-2} \\ \mu_F^2\sigma_F^{-2} + \mu_N\sigma_N^{-2} \end{bmatrix}$$

~~and~~

$$C = \mu_T^2\sigma_T^{-2} + \mu_F^2\sigma_F^{-2} + \mu_N^2\sigma_N^{-2}$$

A1 ~~Likelihood vs Probability~~

Appendix B: Likelihood vs Probability

We note that this ~~is~~ method is distinct from estimating λ as the ratio of the distributions ΔF and ΔT. This is due to a conceptual difference between probability and likelihood. Constructing the likelihood answers the question, "~~(a)~~a: how likely is a particular hypothesis (in this simple case, a particular value of λ) given the evidence?" This is a fundamentally different question from "~~(b)~~b: what is the probability density function of the ratio $-\Delta F/\Delta T$?" The first question involves fixing a putative value of λ, which is *not* treated as a random variable. The second question treats λ as a random variable. Mathematically, this is reflected in the difference between a line integral over the curve $y = -\lambda x$:

$$(a) : P(x, y|\lambda) = \int_C P_{xy}(x, y)ds = \int_{-\infty}^{\infty} P_{xy}(x, -\lambda x)\sqrt{1+\lambda^2}\,dx$$

and the ratio distribution of the random variable $\lambda = -y/x$

$$(b) : P_\lambda(\lambda) = \int_{-\infty}^{\infty} P_{xy}(x, -\lambda x)|x|\,dx$$

We use the ratio distribution ~~(b)~~ b to estimate S once we have the posterior PDF for λ. This is because we treat S as the ratio of two random variables F_{2xCO_2} and λ.

B1 ~~Correlations between $F_{2\times CO_2}$ and ΔF~~

Appendix C: Correlations between $F_{2\times CO_2}$ and ΔF

CO_2 emissions are the primary contributor to present-day radiative forcing change relative to preindustrial. Atmospheric concentrations of CO_2 were lower in the Last Glacial Maximum. This means that the forcing terms ΔF used as evidence in the LGM and historical periods are correlated with the forcing corresponding to doubled CO_2. For visual clarity, we neglect this correlation in this paper. To take it into account, we can write the simple energy balance model as

$$\Delta N = \Delta F' + \beta F_{2\times CO_2} + \lambda \Delta T.$$

In this case, the likelihood $P(E|\lambda, F_{2\times CO_2})$ is defined as the integral of the joint probability distribution of the evidence E over the curve defined by the model. Following S20, we can then calculate S by changing variables and marginalizing over $F_{2\times CO_2}$

$$P(S|E) = \int P(\lambda', F'_{2\times CO_2}|E)\delta(S - F'_{2\times CO_2}/\lambda')(\partial S/\partial \lambda')^{-1}(\partial S/\partial F'_{2\times CO_2})^{-1} dF'_{2\times CO_2} d\lambda'$$

Practically, we can draw samples of λ and $F'_{2\times CO_2}$ from the joint posterior distribution and use these to calculate a posterior distribution for S. This correlation contributes very little to the results; when taking it into account we obtain similar ranges for S as when we neglect it.

Author contributions. TEXT

Competing interests. TEXT

Disclaimer. TEXT

Acknowledgements. ~~KM wishes to acknowledge helpful discussions with Gavin Schmidt.~~ TEXT

References

REFERENCE 1
REFERENCE 2

References

Albani, S. and Mahowald, N. M.: Paleodust insights into dust impacts on climate, Journal of Climate, 32, 7897–7913, 2019.

Albani, S., Mahowald, N., Perry, A., Scanza, R., Zender, C., Heavens, N., Maggi, V., Kok, J., and Otto-Bliesner, B.: Improved dust representation in the Community Atmosphere Model, Journal of Advances in Modeling Earth Systems, 6, 541–570, 2014.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., Webb, M. J., Armour, K. C., Forster, P. M., and Titchner, H.: Accounting for changing temperature patterns increases historical estimates of climate sensitivity, Geophysical Research Letters, 45, 8490–8499, 2018.

Annan, J. and Hargreaves, J. C.: A new global reconstruction of temperature changes at the Last Glacial Maximum, Climate of the Past, 9, 367–376, 2013.

Annan, J. D., Hargreaves, J. C., and Mauritsen, T.: A new global surface temperature reconstruction for the Last Glacial Maximum, Climate of the Past, 18, 1883–1896, 2022.

Armour, K. C., Bitz, C. M., and Roe, G. H.: Time-varying climate sensitivity from regional feedbacks, Journal of Climate, 26, 4518–4534, 2013.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A.-L., et al.: Bounding global aerosol radiative forcing of climate change, Reviews of Geophysics, 58, e2019RG000 660, 2020.

Bereiter, B., Shackleton, S., Baggenstos, D., Kawamura, K., and Severinghaus, J.: Mean global ocean temperatures during the last glacial transition, Nature, 553, 39–44, https://doi.org/10.1038/nature25152, 2018.

Braconnot, P. and Kageyama, M.: Shortwave forcing and feedbacks in Last Glacial Maximum and Mid-Holocene PMIP3 simulations, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 20140 424, 2015.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nature Climate Change, 2, 417–424, https://doi.org/10.1038/nclimate1456, 2012.

Budyko, M. I.: The effect of solar radiation variations on the climate of the Earth, tellus, 21, 611–619, 1969.

Caballero, R. and Huber, M.: State-dependent climate sensitivity in past warm climates and its implications for future climate projections, Proceedings of the National Academy of Sciences, 110, 14 162–14 167, https://doi.org/10.1073/pnas.1303365110, 2013.

Claquin, T., Roelandt, C., Kohfeld, K., Harrison, S., Tegen, I., Prentice, I., Balkanski, Y., Bergametti, G., Hansson, M., Mahowald, N., et al.: Radiative forcing of climate by ice-age atmospheric dust, Climate Dynamics, 20, 193–202, 2003.

Cooper, V. T., Armour, K. C., Hakim, G. J., Tierney, J. E., Osman, M. B., Proistosescu, C., Dong, Y., Burls, N. J., Andrews, T., Amrhein, D. E., Zhu, J., Dong, W., Ming, Y., and Chmielowiec, P.: Last Glacial Maximum pattern effects reduce climate sensitivity estimates, Science Advances, 10, https://doi.org/10.1126/sciadv.adk9461, 2024.

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., Zhou, C., and Andrews, T.: Intermodel spread in the pattern effect and its contribution to climate sensitivity in CMIP5 and CMIP6 models, Journal of Climate, 33, 7755–7775, 2020.

Forster, P. M.: Inference of climate sensitivity from analysis of Earth's energy budget, Annual Review of Earth and Planetary Sciences, 44, 85–106, 2016.

Forster, P., T. S. K. A. W. C. J.-L. D. D. F. D. L. T. M. M. P. M. W. M. W. H. Z.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., P. Z. A. P. S. C.-C. P. S. B. N. C. Y. C. L. G. M. G. M.

H. K. L. E. L. J. M. T. M. T. W. O. Y. R. Y. and Zhou, B., chap. 7, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

Friedrich, T. and Timmermann, A.: Using Late Pleistocene sea surface temperature reconstructions to constrain future greenhouse warming, Earth and Planetary Science Letters, 530, 115 911, https://doi.org/https://doi.org/10.1016/j.epsl.2019.115911, 2020.

Friedrich, T., Timmermann, A., Tigchelaar, M., Elison Timm, O., and Ganopolski, A.: Nonlinear climate sensitivity and its implications for future greenhouse warming, Science Advances, 2, e1501 923, 2016.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian data analysis, Chapman and Hall/CRC, 1995.

Gelman, A., Simpson, D., and Betancourt, M.: The prior can often only be understood in the context of the likelihood, Entropy, 19, 555, 2017.

Gregory, J. M. and Andrews, T.: Variation in climate sensitivity and feedback parameters during the historical period, Geophysical Research Letters, 43, 3911–3920, 2016.

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., and Wagenmakers, E.-J.: A primer on Bayesian model-averaged meta-analysis, Advances in Methods and Practices in Psychological Science, 4, 251524592110312 56, 2021.

Hansen, J. E., Sato, M., Simons, L., Nazarenko, L. S., Sangha, I., Kharecha, P., Zachos, J. C., von Schuckmann, K., Loeb, N. G., Osman, M. B., et al.: Global warming in the pipeline, Oxford Open Climate Change, 3, kgad008, 2023.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Climate Dynamics, 35, 785–806, https://doi.org/10.1007/s00382-009-0630-8, 2009.

Hopcroft, P. O., Valdes, P. J., Woodward, S., and Joshi, M. M.: Last glacial maximum radiative forcing from mineral dust aerosols in an Earth system model, Journal of Geophysical Research: Atmospheres, 120, 8186–8205, 2015.

Kageyama, M., Harrison, S. P., Kapsch, M.-L., Lofverstrom, M., Lora, J. M., Mikolajewicz, U., Sherriff-Tadano, S., Vadsaria, T., Abe-Ouchi, A., Bouttes, N., Chandan, D., Gregoire, L. J., Ivanovic, R. F., Izumi, K., LeGrande, A. N., Lhardy, F., Lohmann, G., Morozova, P. A., Ohgaito, R., Paul, A., Peltier, W. R., Poulsen, C. J., Quiquet, A., Roche, D. M., Shi, X., Tierney, J. E., Valdes, P. J., Volodin, E., and Zhu, J.: The PMIP4 Last Glacial Maximum experiments: preliminary results and comparison with the PMIP3 simulations, Climate of the Past, 17, 1065–1089, https://doi.org/10.5194/cp-17-1065-2021, 2021.

Köhler, P., Bintanja, R., Fischer, H., Joos, F., Knutti, R., Lohmann, G., and Masson-Delmotte, V.: What caused Earth's temperature variations during the last 800,000 years? Data-based evidence on radiative forcing and constraints on climate sensitivity, Quaternary Science Reviews, 29, 129–145, 2010.

Koricheva, J., Gurevitch, J., and Mengersen, K.: Handbook of meta-analysis in ecology and evolution, Princeton University Press, 2013.

Loulergue, L., Schilt, A., Spahni, R., Masson-Delmotte, V., Blunier, T., Lemieux, B., Barnola, J.-M., Raynaud, D., Stocker, T. F., and Chappellaz, J.: Orbital and millennial-scale features of atmospheric CH4 over the past 800,000 years, Nature, 453, 383–386, 2008.

Mahowald, N. M., Yoshioka, M., Collins, W. D., Conley, A. J., Fillmore, D. W., and Coleman, D. B.: Climate response and radiative forcing from mineral aerosols during the last glacial maximum, pre-industrial, current and doubled-carbon dioxide climates, Geophysical Research Letters, 33, 2006.

Marvel, K., Schmidt, G. A., Miller, R. L., and Nazarenko, L. S.: Implications for climate sensitivity from the response to individual forcings, Nature Climate Change, 6, 386, 2016.

Marvel, K., Pincus, R., Schmidt, G. A., and Miller, R. L.: Internal Variability and Disequilibrium Confound Estimates of Climate Sensitivity From Observations, Geophysical Research Letters, 45, 1595–1601, 2018.

Modak, A. and Mauritsen, T.: Better-constrained climate sensitivity when accounting for dataset dependency on pattern effect estimates, Atmospheric Chemistry and Physics, 23, 7535–7549, https://doi.org/10.5194/acp-23-7535-2023, 2023.

Renoult, M., Sagoo, N., Zhu, J., and Mauritsen, T.: Causes of the weak emergent constraint on climate sensitivity at the Last Glacial Maximum, Climate of the Past, 19, 323–356, https://doi.org/10.5194/cp-19-323-2023, 2023.

Rohling, E. J., Marino, G., Foster, G. L., Goodwin, P. A., Von der Heydt, A. S., and Köhler, P.: Comparing climate sensitivity, past and present, Annual Review of Marine Science, 10, 261–288, 2018.

Rose, B. E., Armour, K. C., Battisti, D. S., Feldl, N., and Koll, D. D.: The dependence of transient climate sensitivity and radiative feedbacks on the spatial pattern of ocean heat uptake, Geophysical Research Letters, 41, 1071–1078, 2014.

Sellers, W. D.: A global climatic model based on the energy balance of the earth-atmosphere system, Journal of Applied Meteorology and Climatology, 8, 392–400, 1969.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., and Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484, 49–54, https://doi.org/10.1038/nature10915, 2012.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., et al.: An assessment of Earth's climate sensitivity using multiple lines of evidence, Reviews of Geophysics, 58, e2019RG000 678, 2020.

Siegenthaler, U., Stocker, T. F., Monnin, E., Luthi, D., Schwander, J., Stauffer, B., Raynaud, D., Barnola, J.-M., Fischer, H., Masson-Delmotte, V., et al.: Stable carbon cycle climate relationship during the Late Pleistocene, Science, 310, 1313–1317, 2005.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A.: Bayesian approaches to random-effects meta-analysis: a comparative study, Statistics in medicine, 14, 2685–2699, 1995.

Snyder, C. W.: Evolution of global temperature over the past two million years, Nature, 538, 226–228, https://doi.org/10.1038/nature19798, 2016.

Stap, L. B., Köhler, P., and Lohmann, G.: Including the efficacy of land ice changes in deriving climate sensitivity from paleodata, Earth System Dynamics, 10, 333–345, 2019.

Sutton, A. J. and Abrams, K. R.: Bayesian methods in meta-analysis and evidence synthesis, Statistical methods in medical research, 10, 277–303, 2001.

Takemura, T., Egashira, M., Matsuzawa, K., Ichijo, H., O'ishi, R., and Abe-Ouchi, A.: A simulation of the global distribution and radiative forcing of soil dust aerosols at the Last Glacial Maximum, Atmospheric Chemistry and Physics, 9, 3061–3073, 2009.

Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., and Poulsen, C. J.: Glacial cooling and climate sensitivity revisited, Nature, 584, 569–573, 2020.

von Deimling, T. S., Ganopolski, A., Held, H., and Rahmstorf, S.: How cold was the Last Glacial Maximum?, Geophysical Research Letters, 33, https://doi.org/10.1029/2006gl026484, 2006.

Winton, M., Takahashi, K., and Held, I. M.: Importance of ocean heat uptake efficacy to transient climate change, Journal of Climate, 23, 2333–2344, 2010.

Yue, X., Wang, H., Liao, H., and Jiang, D.: Simulation of the direct radiative effect of mineral dust aerosol on the climate at the Last Glacial Maximum, Journal of Climate, 24, 843–858, 2011.

Zhu, J. and Poulsen, C. J.: Last Glacial Maximum (LGM) climate forcing and ocean dynamical feedback and their implications for estimating climate sensitivity, Climate of the Past, 17, 253–267, 2021.

has not been handled correctly. This leads to a different pdf (e.g. Gaussian + constant). This is a common problem is science - it's called quality control.

Perhaps our approach can be seen as a quality control method? If one expert's prior is much narrower than others, assuming all have access to the same knowledge, then that suggests the first expert is being overconfident.

34.
l.379: "The beauty of Bayesian methods ...". The beauty of objective Bayesian methods is that you don't need to deal in "belief" at all.
We have replaced "beliefs" with "knowledge"

35.
l.396, eq.(A3). Superficially, there appears to be a minus sign missing here – required for
a Gaussian shape.
The integral is correct- it's a commonly used form in quantum field theory. Note that this is the line integral over the curve defined by the model (and that lambda is negative in B).

36.
Equations at the end of section A1. Again, not clear how this leads to a Gaussian shape.
The integrals are correct

Editorial Comments all fixed

37. l.242: typo "is are".

38. l.260, 261.: typos "s(S", "y(K"

39. A2, line 1: typo "this is method is distinct".